THE COMPARABILITY OF STANDARDS SET AT THE ZIMBABWE GENERAL CERTIFICATE OF EDUCATION ORDINARY LEVEL IN GEOGRAPHY 2248 AND INTEGRATED SCIENCE 5006 OVER A PERIOD OF THREE YEARS

BY
LAZARUS NEMBAWARE

A THESIS PRESENTED TO THE FACULTY OF EDUCATION IN FULFILMENT OF THE REQUIREMENT OF THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF SCIENCE AND MATHEMATICS EDUCATION

UNIVERSITY OF ZIMBABWE, HARARE
2004

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

Dedication

I would like to thank my wife, Virginia, for supporting me during the time I secluded myself from the family when I worked on this thesis. I dedicate this thesis to her and our three children, Kudzai, Fungai and Tafadzwa.

Acknowledgements

I wish to express my gratitude to my supervisors, Dr. D.J.K. Mtetwa and Dr. S.M. Chagwedera for their guidance in this study. I also acknowledge the assistance that I got from the Zimbabwe School Examinations Council in the form of the necessary documents that made my study possible.

Abstract

The stakeholders of the Zimbabwean examination system expect a given grade to represent a certain achievement standard despite the year it was gained and the subject in which it was achieved. However, ten years after localising examinations the degree of similarity of the examination standards that were set by the Zimbabwe School Examinations Council within each of the subjects and between any two, at the General Certificate of Education Ordinary level from 1996 to 1998 was still unknown. This study compared the same assessment objectives that are used to test the body of knowledge in Geography 2248 and Integrated Science 5006 from 1996 to 1998.

Examinee-centred and test-centred approaches to comparing examination standards between and within subjects over a period of time were investigated.

The correlational research design was used in the study to find out the relationship between the performances of candidates in two subjects and within each of the two subjects. The survey method was used to collect professional judgements from experts in examinations. The population from which the sample was taken is in strata that stems out of the nature of responsible authorities and geographical locations. The selection of the sample was done through the stratified random technique. A two percent sample out of a population of 110 000 students who sat for both subjects was preferred. Questionnaires were sent out to experts to collect information on their judgements of syllabus content and assessment skills in the question papers offered to candidates from 1996 to 1998.

The study gathered strong evidence to show that the standards set in the two subjects were comparable from one year to the next. However, the standards in Integrated Science 5006 were not as high as those in Geography 2248 indicating that adherence to the setting standards by item writers was not as strict as required by the syllabus. Having found out that there is a similarity in the set standards over the three-year period, a system of reporting achievement is recommended by the study. The findings led to the proposal that candidates' performance be reported using assessment objectives.

This study makes a distinct contribution to the body of knowledge by using a methodology that has not been used before on the Zimbabwe School Examinations Council question papers as well as the comparing, for the first time, of the quality of question papers over time and also the performance of candidates who answered the question papers in two subjects between 1996 and 1998. The frontiers of knowledge are widened by this study because it gives stakeholders of examinations a system of reporting standards embedded in candidates performance qualitatively.

CHAPTER I: THE RESEARCH PROBLEM

Introduction

This chapter looks at the research problem of this study. The research questions that guided the investigation into the comparability of examination standards at the General Certificate of Education Ordinary level, the limitations and delimitation of the study and definitions of terms are also presented.

Statement of the Problem

The purpose of this research was to establish the examination standards set in the subjects of Geography 2248 and Integrated Science 5006 at the General Certificate of Education (GCE) Ordinary ("O") level and investigate whether these standards are comparable from year to year. The comparability of examination standards between subjects and from year to year is what makes public examinations credible to the stakeholders. The users of examination results such as employers and institutions of higher learning expect a given grade to indicate similar standard of candidates' performance irrespective of the subject in which it was achieved and the year that the grade was gained. The stakeholders of public examinations do not know whether or not the assessment objectives that are in the two syllabuses are in fact in each question paper from one year to the next. These assessment objectives are further not used to describe candidates' attainment on the "O" level certificates. In other words, the grades on the "O" level certificates are not described in terms of what candidates know and are able to do. The quality of each grade in the two subjects is, therefore, unknown by the users of "O" level examination results. Furthermore, the

stakeholders do not know the relationship between the grades awarded to the same candidates who wrote Geography 2248 and Integrated Science 5006 examinations at Ordinary level. It was also necessary to establish the validity of the question papers as measuring instruments of candidates' ability.

The localisation of the Ordinary level examination system highlighted the problem of comparability of examination standards. The change from the Cambridge examinations to a local examination system gave the stakeholders the reason to raise the issue of comparability with the new Council. Although the localisation programme brought into the country all the benefits described in Chapter II there wasn't a built-in system to inform stakeholders of the degree of similarity in the standards that were set from one year to the next. This did not mean that the British examinations had a system of informing Zimbabweans on the comparability of standards. Zimbabweans simply accepted that British standards were high because of the long history they had in examinations. The involvement of Zimbabweans in the examination system, in a way, alerted them to the issue of comparability. It was one of the fundamental concerns of this study to provide answers to questions on comparability of examination standards within subjects and between subjects. Having worked for the Examinations Branch of the Ministry of Education which was later transformed into the Zimbabwe School Examinations Council from the time syllabuses were developed up to the time of this study, the need to empirically show the stakeholders the degree of similarity between standards that were set was motivational to this researcher.

Research Questions

Three research questions guided investigations into the research problem referred to above. These are stated below.

1. Were the question papers in the two subjects set according to the assessment objectives outlined in the syllabuses?

2. Were the Geography 2248 and Integrated Science 5006 question papers valid instruments for measuring the content outlined in the syllabuses?

3. What was the nature of the relationship between the grades that were awarded to the candidates within each of the subjects over the three-year period and also between the same candidates who wrote the two subjects in each of the three years?

Assessment objectives are indicators of standards in examinations in that they are the skills that are tested by examination question papers. Each of the syllabuses has a number of assessment objectives that must be included in question papers. Determining whether or not the question papers were set according to the assessment objectives will help us to establish the degree of similarity of standards over the three-year period. Each of the syllabuses in Geography 2248 and Integrated Science 5006 have assessment objectives outlined so that teachers, students and examiners are aware of the structure of papers in as far as the skills that are examined.

The second research question logically follows the first one in that the syllabus content that must be in question papers is upon which candidates should demonstrate the skills mentioned in the first research question. It was important to investigate the

consistency of the content in the question papers over the three-year period and the extent they reflected the domains used by teachers in preparing their candidates for the examinations. The domains were clearly stated and described in the syllabus documents.

The strength of the relationship between the grades that are awarded to candidates from one year to the next has a strong bearing on comparability of standards. This is a relevant question as stakeholders of the education system expect a given grade in one year to indicate similar ability level of candidates in another year. A strong positive relationship would mean high comparability of standards from one year to the next and from one subject to the other while a weak or negative relationship would mean that there is a slight or no relationship between the grades awarded to candidates.

Importance of the Study

Given that the Zimbabwean society places so much emphasis on passing examinations, this study should be of significance to many people. It is important for the stakeholders in the Zimbabwean education system, industry and commerce, institutions of higher learning and parents to know whether or not the question papers are reflective of the statements that describe the quality of the question papers as stated in the syllabuses. The Standards Control Unit and the Curriculum Development Unit, both of the Ministry of Education, should also find this study significant, as it will lay down the basis of comparing standards within a subject and across subjects. These two units are in charge of quality control in the Zimbabwean education system. Providing qualitative statements, which depict performance,

should impact on the teaching of students because teachers will know the type of performance expected at each “O” level grade. The knowledge of the expected performance standard or criterion by teachers would have a very significant wash back effect on the way teachers discharge their duties. Teachers are likely to adopt this approach of reporting performance during the mid-year examinations because students would know the areas in which they have strengths and weaknesses. The study will also be of significance to the Zimbabwe School Examinations Council (ZIMSEC) in assisting in the evaluation of the grading system of "O" level candidates. In particular, it will help the organisation in providing statements which describe performance at grades A to U. The benefit of this approach to everyone will be that examination results will be understood in terms of knowing what each holder of a certificate knows and is able to do.

The research problem was looked at in the context of the operations of the Zimbabwe School Examinations Council, the parastatal that has the responsibility of running examinations at all levels in the primary and secondary education system in Zimbabwe. The Council has the sole responsibility of setting and maintaining examination standards in the country. I should hasten to add that the issue of setting and maintaining standards is not for the Zimbabwe School Examinations Council alone. All examination boards in the world have to address this issue or run the risk of having stakeholders rejecting the examinations that they offer. By 1996 (the year which this study refers to) the Zimbabwe School Examinations Council had been running the “O” level examinations for six years. In this regard, it was necessary to give the historical background of the localisation of the Ordinary level examinations

so that the reader could understand where the benchmark to the Zimbabwe School Examinations Council standards came from. This area is covered in Chapter II.

Limitations of the Study

The study used experts in the evaluation of question papers that were set in Geography 2248 and Integrated Science 5006. The number of these experts was limited. An expert in any of the two subjects analysed in this study was one who would have taught the subject at Ordinary level, set national examinations and participated in the marking of the subject. They would have further been involved in the grading of candidates and the reviewing of candidates' grades. Participation in the setting and marking of the subject meant that an individual would have gone through a training process by the Zimbabwe School Examinations Council. There were not many experts in the two subjects and this is because the Zimbabwe School Examinations Council does not use too many people in the setting of each paper in a syllabus. Though the study would have been better off with more experts only seven in each subject participated.

It was not possible to use marked scripts in the two subjects, as this would have compromised the security that is put in place by the Zimbabwe School Examinations Council. The Council's policy is not to let members of the public use such material for research. The use of scripts would have helped in the investigation of how cut-off scores are translated into grades and so give further ground for the analysis of whether or not that cut-off score was used from one year to the next. The emphasis was, however, placed on the question papers that must carry the standard that is set in

the syllabus. How candidates responded to the standard set in the question papers was evaluated by grades that were awarded to the candidates over a period of three years.

Definition of Terms

Standards are levels of performance of candidates that are considered appropriate and adequate at certain stages in our education system.

A Syllabus is made up of the content which students must learn over a period of time in order to adequately prepare for an appropriate examination. A syllabus grade or mark is taken to mean the aggregated or weighted mark from the papers that are examined in a syllabus.

Assessment Objectives are indicators of the performance of candidates who take an examination. The objectives are the skills which candidates are expected to demonstrate at different levels of attainment.

Geography 2248 is the syllabus in the subject of geography that is offered to candidates by the Zimbabwe School Examinations Council. The code 2248 differentiates the syllabus from any other geography syllabus.

Integrated Science 5006 is the syllabus in the subject of Integrated Science that is offered to candidates by the Zimbabwe School Examinations Council. The code 5006 differentiates the syllabus from any other Integrated Science syllabus.

Weighting of assessment objectives is the percentage of marks allocated to each assessment objective tested by a question paper.

Specification grid is a matrix that shows the content, assessment objectives and marks which each question paper in a syllabus would have.

Candidates are the students who wrote the Geography 2248 and Integrated Science 5006 or any other subject that is administrated by an examining authority.

Content validity is when a question paper tests the syllabus areas it is expected to test.

An Examiner is a person who sets examination question papers

Delimitations

This study was confined to the comparability of standards that were set in the ordinary level Geography 2248 and Integrated Science 5006 question papers over a three-year period, 1996 to 1998. The standards that were investigated were those that were test and examinee centred. It was not the intention of the study to compare standards that were set by Zimbabwe School Examinations Council and the University of Cambridge Local Examinations Syndicate as the study was not to compare standards of different boards that used different syllabuses. The comparison was on whether or not question papers set by ZIMSEC had the same skills and content, and whether or not the pass rate was reflective of the standards set by the question papers. The study then went on to link the skills that were found in the question papers to a way of reporting performance of candidates. The study used a random stratified sample taken from the population that wrote the two subjects. Consequently the results should be representative of the developments in the two subjects at the General Certificate of Education Ordinary level.

Summary

This chapter set out the research problem for this study and this was that the stakeholders of public examinations do not know whether or not the assessment objectives that were in the two syllabuses were in fact in each question paper from one year to the next. These assessment objectives were further not used to describe candidates' attainment on the "O" level certificates. The research questions were stated in this chapter because they directed the investigations into the research problem. The importance, limitations, and delimitations of the study were also discussed in this chapter.

CHAPTER II: CONCEPTUAL FRAMEWORK

Introduction

This chapter focuses on the review of relevant literature in the area of setting examination standards, interpretation of test scores and the comparability of standards. The historical background of the localisation of the "O" level examination system is also provided in this chapter. The conceptual framework of the study is set as these issues are looked at in this chapter. Standards in public examinations have to be set, accepted by most of the stakeholders and then maintained at an acceptable level in order for certificates awarded by an examination board to have credibility. The inconsistency in setting standards has ripple effects on the maintenance and comparability of examination standards. Before discussing the issue of comparability of standards, it is important, to provide some vital information on the formation of ZIMSEC so that the reader appreciates how the issue of standards is linked to the establishment of the Council that runs examinations in Zimbabwe. It has been pointed out in Chapter I that the stakeholders of examinations in Zimbabwe have no basis of comparing standards in terms of what students know and are able to do. The alpha grades which a candidate is awarded fails to explain to a stakeholder the actual level of what a candidate can accomplish if given a task to do. Users of examination certificates only know that grade A is better than grade B, and grade B is better than grade C and so forth, but comparisons without the qualitative comments provide a weakness for stakeholders. This study rests on the ideas that are propounded by Matthews (1985), that if a large group of candidates write examinations in two subjects that have similar standards, the mean grades obtained

by candidates in the two subjects would be the same. Lack of similarity means that standards are not the same. Comparing standards using the mean grades only does not address the problem which this study seeks to address. The approach used by Matthews (1985) has a weakness in that only one aspect of examinations, the results, is compared. In this study I have worked on three critical elements in examinations standards. This makes the comparisons made in this study better than the methods used by Matthews (1985). In this study I link assessment objectives as discussed in the objectives movement by Kelly (1999) and performance of candidates. This study focuses on how equivalent standards are in each of the subjects from year to year and also between them using the test centred and the examinee centred approaches to standards setting.

Formation of the Zimbabwe School Examinations Council

Overseas Examination Boards

Three overseas examination boards controlled the setting and marking of Ordinary level examinations in Zimbabwe before 1984. The boards were the University of Cambridge Local Examinations Syndicate (UCLES), the Associated Examinations Board (AEB), and the University of London School Examinations Board (ULSEB). It is important to note that examinations in the local languages, Ndebele and Shona, were even before 1984, set and marked by local examiners and markers. The external examinations boards ran their examinations through the Ministry of Education. The Ministry had a section called the Examinations Branch that implemented the ministry's policy on examinations. The role of the Examinations Branch was, before the localisation programme, an administrative one in that it

received syllabuses and question papers from overseas and distributed these to examination centres throughout the country. Upon completion of examinations, the scripts were freighted overseas for marking. The overseas examination boards processed the marks and released the results for the candidates through the Examinations Branch. It is important to note that Zimbabweans were not responsible for quality control issues such as comparability of standards, reliability of marking and validity of question papers in the examinations before the localisation of examinations. The overseas boards were responsible for these. However, the Examinations Branch had the responsibility to ensure that examinations were written according to the rules and regulations of the overseas boards. Though the overseas boards were running a business, the Examinations Branch was providing a service to students because it was not paid for carrying out the administrative work for the boards. The Examinations Branch collected examination fees on behalf of the overseas boards and remitted the money to the boards. The use of overseas examinations boards was a weakness for Zimbabweans in the area on comparability of examination standards as there was no involvement in this area. Not only were examinations offered by many boards but there were based on different syllabuses.

Syllabus Development

The government tasked the Ministry of Education with the responsibility of ensuring that the examination system was localised. Two sections of the Ministry of Education, the Examinations Branch and the Curriculum Development Unit (CDU), worked closely together to ensure that the set targets in the new programme were met. The development of syllabuses is the logical starting point for the localisation of an examination system in that syllabuses are the documents that have the

information on content and skills which students must learn before they sit for an examination. This was the reason for the CDU and the Examinations Branch to work together.

The Geography and Science syllabuses that were developed by the British examination boards did not state the assessment objectives which candidates were expected to demonstrate in examinations. An example of syllabuses that were used from the University of Cambridge Local Examinations Syndicate is given as Appendix A and Appendix B. These syllabuses only indicated the content that candidates were expected to cover before they wrote an examination. The instructions to students and teachers were not as explicit as they are now. For example, the University of Cambridge Local Examinations Syndicate Geography 2222 syllabus stated that candidates will be required to answer two questions from section A and two questions from either section B or section C but it did not go further to indicate specifically where the questions were to come from. It, however, states that Section A would have five question set on Malawi, Zimbabwe and Zambia. The syllabus could have gone further to indicate the number of questions on each country or on a geographical theme in those countries. The Geography 2248 and Integrated Science 5006 syllabuses that were developed by Zimbabweans clearly stated the assessment objectives (see Appendices C and D) and the number of questions that would be set on a specific area. It can be noticed that the two Zimbabwean syllabuses have assessment objectives of knowledge with understanding, application of skills, judgement or decision-making, and experimental

skills. Furthermore, the syllabuses show the weighting of the assessment objectives. It is now possible to measure candidates' performance against levels of performance as defined by the assessment objectives. It is also possible to check whether the questions set in a particular subject adhere to the assessment objectives indicated in the syllabus. The objective of the localisation of the syllabuses was to develop syllabuses that were consistent with the policies and national goals of Zimbabwe. The British syllabuses that had been used in the schools in Zimbabwe were designed for international students and a few aspects of the syllabuses were given some local flavour. This was why Zimbabwean pupils studied in some depth the Geography of Central Africa and Canada, English Literature and not Literature in the English language, etc. However, many of the domains in the British syllabuses that were studied by Zimbabwean students were not Zimbabwean. The localisation programme motivated this researcher to investigate the similarity of examination standards or the lack of it. Two aspects of the syllabuses are looked at in this study. These are the content and assessment objectives.

Panels of examiners that were involved in the development of the syllabuses included Education Officers, practising teachers, and those groups with interests in secondary school education. They made recommendations to the Ministry of Education on the content which they wanted incorporated in the syllabuses. After the approval of the new Zimbabwean syllabuses, work on the development of the first examination in Science and Geography started in 1987 and1988 respectively. Consultants from UCLES together with Officers in the Test Development and

Research Unit of the Examinations Branch conducted question-setting workshops. The Examinations Branch sent some of its Officers to spend three months on an attachment to UCLES. The Officers got skills of question paper setting, syllabus design and administration of examinations. Having been trained by the personnel at the University of Cambridge Local Examinations Syndicate in the United Kingdom, this researcher became one of the local people who teamed up with consultants from the UK in training the new Zimbabwean question setters. This first training session in question setting was in 1988. It became necessary to give the new examiners skills of setting question papers since they were doing it for the first time. The objective of the workshops was to develop question-setting skills and also to boost confidence in those who had some knowledge of assessing candidates' performance. Specification grids were used to ensure that the questions produced strictly adhered to the demands of the syllabuses. The Science and the Geography examinations that were based on Zimbabwean syllabuses were written for the first time in 1989 and 1990 respectively. In January 1991 the first Grade and Grade Review sessions were held at the Examinations Branch in Harare under the assistance of an UCLES consultant. These grade and grade review sessions were for the November 1990 examinations. A grading session is where the grade thresholds are determined while a grade review session is where the work of candidates that is on grade borderlines is reviewed by a panel to determine whether or not the grades awarded were indeed the ones which the candidates deserve. This is just a second check on the quality of marking scripts so that an awarding Board is sure that candidates are awarded grades which reflect the work presented. The production of the computer data for the

grading of the candidates, the review of their work and the scaling of marks were done by the Computer Unit at the Examinations Branch for the first time in June 1995.

Candidate Entries

The choice of the board to spearhead the localisation programme was based on the amount of experience the board had in the localisation of examination systems and the number of candidates that was taking examinations with that board. UCLES had, among other boards in the UK, the highest entry of "O" level examination candidates from Zimbabwe. For example, in 1983, 19 023 candidates were offered examinations by UCLES while 4 494 and 11 260 wrote examinations offered by the AEB and ULSEB respectively (Secretary of Education Annual Report, 1984). ULSEB catered for the adult candidates and so the entry was not from formal schools. The candidate entry for the UCLES examinations rose as a result of the phenomenal rise in the enrolment of pupils in schools following the democratisation of access to education in 1980. Masango and Nembaware (1991) note that the system of education before Zimbabwe's independence in 1980 had been designed to allow a small percentage of primary pupils into secondary education. This changed in 1980 and, therefore, led to the rise in secondary school pupils. The Zimbabwe Junior Certificate (ZJC) examination written after two years of secondary education further reduced the number of candidates who proceeded to write the "O" level examinations. Consequently only a small percentage wrote the "O" level examination. Table 1. below illustrates this point. It can be noticed that by 1992,

twelve years after the democratisation of access to education, the number of candidates who were acquiring "O" level education had risen phenomenally.

Table 1. Candidate Entries

YEAR	GRADE SEVEN CANDIDATES	ZJC CANDIDATES	"O" level CANDIDATES
1978	81 903	17 485	13 168
1979	82 210	16 031	12 201
1992	275 557	157 461	218 497

Source: Examinations Branch & ZIMSEC files

The Government of Zimbabwe had to ensure that all schools in the country supported the localised examination system. Section 63 of the Education Act enforced this. It states that "The secretary shall determine curricula and examination system for all schools and in so doing shall not determine different curricula and examination systems for different schools on the grounds that they are government or non-government school." (p.628).

This is significant in that many schools could have opted to continue with the British syllabuses and examinations thereby posing problems of viability for the local Examinations Board. The local Board, being new, would also not be taken seriously by schools that could, because of tradition, view it suspiciously.

The processing of all the entries for the Zimbabwean syllabuses by the Examinations Branch started in 1990. The taking up of examination responsibilities from UCLES was a process over a long period of time. This was deliberate and was based on

sound judgement that once an aspect of examinations had been taken over, the personnel in Zimbabwe needed to fully understand it before taking up new responsibilities.

Reasons for the Localisation of the “O” level examinations

By the time the Zimbabwe School Examinations Council Act was passed by the parliament of Zimbabwe in 1994 the localisation programme was almost coming to its fruition. It became necessary for the localised examination system to be run by an autonomous body, the Zimbabwe School Examinations Council, hence the formation of the parastatal. The localisation of examinations was necessary because of educational, economic, and political reasons. These reasons are discussed below.

The British Examinations Boards had made the "O" level qualification internationally recognised. An arm of the Zimbabwe government would not give the examination the same international credibility as an autonomous body would due to possible fears by the users of examination results, that politicians would bring pressure to bear on civil servants to act unprofessionally when processing the examination results. It was, therefore, imperative that a Council that was independent of government is made the custodian of the examination standards already established by the British Examinations Boards.

There was need to provide teachers in the school system with feedback on the performance of students in national examinations. This feedback helps in improving

curriculum delivery to students. Teachers were to have the opportunity to be involved in the marking of candidates' scripts. This gave those who were involved the skills of marking their own students' school based assessments professionally.

The provision of "O" level examinations by foreign examination boards involved colossal amounts of money in foreign currency. The Zimbabwe dollar was not stable against the British pound and as a result the amount to run national examinations continued to rise every year. The rise in the entry of candidates each year also meant a rise in the amount of money remitted to UCLES. For instance, in 1989 ZW$16 358 544,92 was remitted to UCLES but by 1992 the amount had gone up to ZW$44 113 632,66 (Deputy Minister of Education Speech at UCLES, 1996). It was envisaged that the localisation of the examination system would effectively cut down this rise in the cost of examinations.

Zimbabwe became a sovereign state in 1980. It was now time that an independent state ran its own examinations rather than leave it in the hands of the former colonial power.

The Establishment of the Examinations Council

The Zimbabwe Cabinet approved the Ministry of Education's proposal to localise the "O" level examinations under the supervision of UCLES in August 1983 (Kachale, 1983). The choice of UCLES was based on the information gathered by Maraire

(1982) who held meetings with the officials at UCLES, AEB and ULSEB, the three boards that were already offering examinations in Zimbabwe.

The Ministry of Education made extensive research both within and outside the country in its efforts to establish an autonomous body that was tasked with the running of examinations in Zimbabwe. Many reports were presented to the Ministry of Education between 1982 and 1986. These were Maraire's Report (1982), Frank Wild's Report (1984), Tanyongana's Report (1985), Mukhurazhizha and Masango's Report (1985), Brearley's Report (1986), and the Ministry of Education Task Force Report (1986). These reports show that a lot of groundwork was done before the passing of the Zimbabwe School Examinations Council Act by parliament in 1994. The reports drew experiences and comments from the United Kingdom, Tanzania, Malaysia, Singapore and Zimbabwe. The Overseas Development Agency (ODA), an arm of the British government and the United States Agency for International Development (USAID) of the United States of America took part in the localisation of examinations. The ODA and USAID funded the consultancy that the University of Cambridge Local Examinations Syndicate provided and the training programmes of the Examinations Branch personnel and markers.

In order to understand the amount of work that went into the preparation for the Council it is necessary to describe some of the reports mentioned above. The reports underlined the importance of examinations standards.

In September 1984, the secretary of UCLES produced a report that gave the Ministry of Education four options for the establishment of an examinations Council. This was the first report after the Ministry of Education had been given the go ahead by the Zimbabwe Cabinet in August 1983 to localise the “O” level examinations. There were four options from which the Ministry of Education could make a choice.

The first option was that the examinations Board would be an advisory one to the Ministry of Education. This meant that the Examinations Branch would stay an integral part of the Ministry of Education and then take the administrative role of examinations.

The second option was that the Board would be an autonomous and self-governing body that is established by a statutory instrument. The body would receive policy from the Ministry of Education.

The third option was that the Board would be established by a statutory instrument but be responsible for the "O" and "A" level examinations only. The Ministry of Education would run the Grade Seven and the Zimbabwe Junior Certificate examinations.

The fourth option was that it would become an integral part of the University of Zimbabwe. This would be the same as the University of Cambridge Local Examinations Syndicate and the Examinations Board in Singapore.

The Ministry of Education studied the Wild (1984) options but could not make any follow up discussions with Wild because he died immediately after presenting his report. The Tanyongana Report was based on a trip to the UK between May 25 to

June 3 and 12-30 June 1985. This report was a follow up to the recommendations that Wild (1984) had presented to the Ministry of Education on the establishment of an examinations council. Since Wild had died immediately after presenting his report it became necessary for the ministry to get first hand information on some of his recommendations from the boards in England. The Tanyongana report was very comprehensive. Before producing the report he consulted UCLES on their organisational structure, the relationship of UCLES with the University of Cambridge, skills needed in examination administration, the services that UCLES received from outside its institution, its spatial requirements, and the merits and demerits of Wild's options. He also gathered information of how the AEB and ULSEB were being run. This report is evidence to wide consultation on which option the Ministry of Education was to take when establishing the Examinations Board. The report recommended that the Ministry establish the Examination Board as an autonomous and self-governing body which would receive policy from the Ministry of Education.

The reasons for choosing this option are discussed in the report. The first reason was that the educational standards could be measured without any risk of pressure from the government. The credibility of examinations already achieved by the overseas boards that were operating in Zimbabwe would be maintained. It was absolutely important to ensure that the stakeholders of examinations become confident that the localisation of examinations did not mean the lowering of educational standards.

The second reason was that there could be a risk of pressure being exerted to alter examinations in accordance with the administrative requirements rather than genuine educational reasons if the examinations Board was to be part of the Ministry of Education.

The third reason was that an autonomous examinations Board could easily recruit specialist staff to run examinations. Recruitment of such staff would be difficult if the Board is run by an arm of government which tends to be bureaucratic. An examination system needs computer specialists to run the electronic processing of candidate entries and results, and highly qualified question paper developers to produce papers of great repute

The report also proposed an organisational structure for the examinations council, the qualifications expected of the person who would chair the Board, the composition of the Board, the Board's committees and the qualifications of the director of the Council. Recommendations on building, furniture, and printing press requirements, services such as postal, transport and cleaning were also presented.

On 18 November 1985 the Minister of Education, Dzingai Mutumbuka appointed a 12-member task force on the localisation of examinations (Chouhan, 1983). It was composed of the Deputy Chief Education Officer for Examinations, the Statistician at the Examinations Branch, the Chief Education Officer at the Curriculum Development Unit, Planning Officer from the Ministry of Education Head Office, Deputy Chief Education Officer for Teacher Education, Chief Executive Officer, Senior Educational Psychologist, two representatives from the University of Zimbabwe, and one person each from UCLES and the British Council. It can be noted that the people who made up the task force were from the relevant sections of the Ministry of Education. The representatives from the University of Zimbabwe ensured that the ideas from the then only university in Zimbabwe were known.

The task force was given three terms of reference. These were to make recommendations on the establishment of the Zimbabwe Examinations Council;

draw up the functions and powers of the Council; and draw up a full localisation programme.

The task force met eight times before presenting a report to the minister (Chouhan, 1983).

A consultant from UCLES, presented his report on the needs of a computer system in 1986 (Brearley, 1986). The report pointed out that the Council would need a computer at a cost of ZW$1,5 million. The report points out that computers already being used for the Grade Seven and Zimbabwe Junior Certificate examinations had inefficient and overloaded hardware and software. The report basically rejected the adaptation of the USA developed computer examination system for three reasons. First, the report pointed out that the system was not well designed. The second reason was that the system had too many programmes that had been poorly written and inadequately tested. Lastly, Brearley rejected the development of the existing computer system because the programmes had not been well documented. He recommended that a new computer be bought and new programmes developed. The programmes were to be the same as those UCLES was using to enable a smooth take over. He recommended that from 1988 the take-over from UCLES start. The stages of the take over were the entry registration, results, certificates and finally examination statistics. During the take-over magnetic tapes were to be passed to and from UCLES. It was envisaged in the report that the full take over would be in 1990. The Brearley report also made recommendations on the personnel for the computer unit of the new council and the floor space that was needed by the computer.

With such consultation and research on how other examination boards had localised examination systems and were running their own examinations it can be concluded

that the architects of the Zimbabwe School Examinations Council Act were provided with enough information upon which to base their decisions. The Act put in place a sixteen-member board that was given the responsibility of running the Council. The board was responsible for policy while the day-to-day issues of the Council were the responsibility of the Director. The chairman of the board at the time of this study was the vice chancellor of the National University of Science and Technology (NUST) and the vice-chairman was the pro-vice chancellor of the University of Zimbabwe. The Council at the time of this study had five divisions. These were the examinations administration, human resources, finance, information services and the test development, research and evaluation. Some professionals in the education system were contracted by the Council to set question papers and mark candidates' scripts. In all, the Council employed over nine thousand contract staff at the peak period. The personnel at the Council had the responsibility of ensuring that the standards in examinations being investigated in this study were set and maintained.

The issue of maintaining examination standards is embedded in the Zimbabwe School Examinations Council Act. Section 4 (1) (h) of the Act stipulates the functions of the Council. One of these functions is " to do all things necessary to maintain the integrity of the system of examinations in respect of primary and secondary education in Zimbabwe." (p.69).

It was, therefore, clear in the minds of the architects of the ZIMSEC Act that integrity of the system of examinations was vital to the very existence of an examinations board, particularly to the one which was starting to establish itself. The maintenance of the integrity of the system of examinations was, in this study, taken to mean the standards that were set in the question papers, the quality of marking candidates' scripts and grades given to candidates. It must be understood that this

function could also refer to many other aspects of examinations such as administration and processing. However, these are not the focus of this study.

The first phase of the localisation of the Zimbabwe General Certificate of Education examinations started in 1984 with the marking of six papers by one hundred and nineteen markers who had been trained by eight subject specialists from the University of Cambridge Local Examinations Syndicate (Secretary of Education Report, 1985). The number of markers and the components handled in Zimbabwe rose remarkably between 1984 and 1996. Table 2 illustrates the rise in the numbers. It was obvious that as the number of markers and components that were being marked by Zimbabweans rose, there were some implications on examination standards. Those in the leadership of marking candidates' scripts had to ensure that inter-marker reliability coefficients during the marking had to be kept very high.

Table 2. Number of Markers and Syllabus Components

YEAR	MARKERS	COMPONENTS
1984	119	6
1985	354	11
1986	1 301	17
1997	4 716	89

Source: Examinations Branch & ZIMSEC files

The localisation programme led to the procurement of machinery such as photocopiers and stand-alone computers. Co-ordination meetings demanded that the scripts that were used be photocopied to facilitate the marking of the same scripts by the markers during the coordination of marking. This was done before markers were given live candidates' scripts. With so much in terms of equipment and human

resources being ploughed into examinations, it became important to gauge whether or not ZIMSEC was getting the deserved return on investment in as far as the maintenance of examination standards was concerned. The information provided in this study on the establishment of a Board to run examinations in Zimbabwe illustrates that the issue of high standards in examinations was a very serious matter that demanded the setting up of a Board that would take on from where the British Boards left. Lack of such an institution would have meant lack of credibility in the localised examination system. The developments in the syllabuses that were used in this study show a great departure from those used on the old syllabuses. Standards were not only being communicated to stakeholders through the syllabuses in terms of content but also through the assessment objectives.

Standards Setting Methods

It is important for a study of this nature to pinpoint what is meant by standards because the concept of standards has been interpreted in a variety of ways in different education systems at any particular time (Goldstein & Heath, 2000). The word standard has been used in education in three distinct ways (Gipps, 1990). The three ways are attainment, levels of educational provision, and matters of conduct and social behaviour. The first use refers to that which a student acquires or accomplishes after studying a course. That attainment is in the area of the educational objectives or criteria that is set out in the syllabuses. The levels of educational provision refer to materials that are given to pupils, teachers, and the school in order to achieve what they are intended to. If students were undisciplined then the standard of their social behaviour would be unbecoming of them. That is

the third use that is mentioned above. As indicated in Chapter I, it is the first use of the word standard which this thesis focuses on.

Standards are defined by Satterly (1981) as the attributes of performances held up as representations against which judgements can be made. This refers to what the student would have attained or that which they are expected to have attained after undergoing a course. Here, Satterly (1981) is in agreement with Gipps (1990) that a standard is that which is attained by candidates. Satterly (1981) points out that a standard " would be a model of attainment or performance used for comparison or measurement, or as a target held up to learners as a basis for their aspirations." p. 251.

So a standard can be seen as a criterion that examinees need to achieve in order to be deemed competent. These criteria can be in the form of the statements of attainment. A statement of attainment describes candidates' expected behaviour after having gone through a course. The syllabuses used by candidates who prepared for the Geography 2248 and Integrated Science 5006 had these expected behaviours listed as assessment objectives or the required skills. The extent to which the candidates can demonstrate capabilities of any competency is upon which judgement can be passed. Judgements on the performance of candidates must be made against how candidates' answers relate to the skills reflected in the questions of a particular test. In other words, the assessment objectives that are in the syllabuses are the benchmarks for judgements on candidates' performance. Examination standards are, therefore, the demands of syllabuses and their assessment arrangements and the levels of performance which candidates must achieve to be awarded certain grades. There is yet another interpretation that can be given to the word standard coming out of Satterly's (1981) definition. If a standard is 'performance used for comparison' it can be argued that norms that are used in norm referencing when interpreting test scores are a standard

as much as the use of criteria described above. The point is standards can be set using different methods.

In defining examination standards, Cresswell (1996) points out that there are critical areas on which value judgements must be made. The areas are: ***what*** attainments are assessed and the ***quality*** of the observed performance during which those attainments are demonstrated. I have addressed these two areas in this study in that the ***what*** is the content upon which judgements are passed by experts and the ***quality*** are the skills which the question papers were testing. Cresswell (1996) points out that defining examination standards must always involve a consideration of the value attached to the knowledge, skills and understandings which are assessed.

This point is important in that no one can define standards of anything he/she does not know about. Therefore, the determination of comparable standards in examinations must involve people who are experts in that field.

Having understood what standards are, it is now possible to move on to the area of standards setting.

There are many methods of setting educational standards in public examinations (Crocker & Algina, 1986). These methods are based on human judgements and as a result there could be as many standards as there are judges if what is deemed the standard is not clearly stated. The Task Group on Assessment and Testing Report (DES, 1989) warned that in the absence of powerful external evidence to show the level of pupils' achievement in school based assessments, teachers' expectation become

teachers' standards. The report argued that it is important to validate school-based assessments with valid and reliable external tests to minimise the subjective judgements of teachers. This highlighted the problem in standards setting. Rowntree (1987) warns that "even within a subject, the standards being maintained are more probably assessment procedures rather than standard attainments." (p.21).

This point cannot be true where the expected standard is embedded in the questions that examinees are expected to answer. The questions would have the assessment objectives which are the attributes of performance upon which those people who are concerned with standards can make judgements. In other words, elaborate syllabuses that have benchmarks for standards ensure that standards thus referred to are not mere assessment procedures but models of performance. Moreover, when the job of ensuring that standards are set and maintained is being done as is expected, there would be blue prints that would be used to check what one purports to have done when setting standards. Assessments procedures can, therefore, not be confused with standard attainments.

I have, therefore, identified the assessment objectives in the syllabuses of Geography and Integrated Science which must be used as a basis of checking whether or not they are in the question papers that were set from 1996 to 1998. Their presence or absence in the question papers will provide a platform to judge the quality of the examinations that were offered to candidates that sat for these examinations.

Weirsma and Jurs (1990) list five methods of standards setting. These are: the professional judgement, Nedelsky's method, Angoff's method, contrasting groups, and standards from norms.

The professional judgement method of determining standards depends solely on the judgements of the experts used. There must be objectives that are used as criteria against which the judgement of a particular standard is made. If objectives were not used there would be a problem of subjectivity. Use of this method alone in examinations could be disastrous as standards could change as the people used change.

Nedelsky proposed his method of setting standards in 1954 (Weirsma & Jurs, 1990). The technique which is used to set standards in multiple choice tests requires that people who are knowledgeable in the content that is tested set the standards. A panel of judges determines the minimally acceptable competency on a test to be administered. Each of the panel members indicates the options on each test item that a minimally competent candidate can eliminate. If, for example, a candidate can eliminate one option in a four-option item then the probability of guessing is 1/4. The probabilities of all the judges are added up and an average is calculated. That average becomes the standard for the minimally competent candidates.

Angoff's method also uses a panel of judges. The experts examine each item and estimate the percentage in a group of minimally competent persons who could answer the item correctly. The minimally acceptable score from each expert is added up and an average found. The average becomes the standard of minimally acceptable performance.

Contrasting group also uses a panel of judges. Two groups of examinees, one that is judged as having mastered the content taught and the other that has not, are used. The test scores of the two groups are plotted on a graph and the point where they intersect becomes the score that indicates minimally acceptable performance.

The last method mentioned above, the use of norms, means that normative performance from known groups is used to set standards.

The methods described above involve the passing of decisions on test content; the difficulty level of items which constitute a test; and also looking at the performance of the candidates. It has been shown that the methods of standard setting are diverse. Shepard (1984) points out that there is scope in using both the test content approach and the examinee performance approach. Shepard proposes that as much information as possible be collected before a standard is set. To this end normative data should be used in conjunction with criterion - referenced approach in order to get as much information as possible.

This emphasizes the fact that in order for standards setters to reach informed decisions, as much information as possible must be at their disposal and so norm referencing and criterion referencing of test scores must be looked at in a complementary way.

Four out of the five methods described above make use of experts in setting standards. The method used by the Zimbabwe School Examinations Council is the professional judgement method of determining standards. As explained above, a panel of experts who in this case are the ZIMSEC's Subject Managers and the Chief Examiners pass judgements on what it is a candidate should have demonstrated to earn a certain grade. A panel of judges that is made up of markers and ZIMSEC Officials make grading decisions that are based on the calibre of candidates in a current examination and that of the previous examination; and the quality of the question paper and how it compares to the one written in the previous year.

The judges used by ZIMSEC are experts in the content students would have studied before writing an examination. Appendix E shows the form which a chief examiner must fill in before the standards fixing meeting. This is called the Grade Threshold Recommendation Form. Chief examiners must respond to four areas on the form. They must fill in the syllabus code, the component, and the maximum marks of the component. The Chief Examiners have to respond to questions that have a strong bearing on examination standards. These questions are on whether or not the current examination component was of the same difficulty, more difficult or less difficult than that of the previous examination; and whether or not there was any reason to think that the quality of candidates was the same as or better or worse than that of the previous year.

Chief Examiners have also to suggest cut-off points from the point of view of the answers presented by candidates. The chief examiners are given the previous examination scripts at grades A, C and E so that they can compare the standard of answers before they make a recommendation. Statistical information on the performance of the candidates in the form of the distribution of marks, paper and syllabus mean marks and standard deviations are also used. The technique of setting standards described here shows that ZIMSEC does not only use a team of judges who only look at the qualitative aspects of grading candidates but also the statistical information. Clearly normative data and criteria are used to arrive at a standard at the Zimbabwe School Examinations Council. A weakness of this system is that there is a lot of evidence that is taken from the candidates and not so much from the question papers. There is no reference to the assessment objectives that were used at question paper setting stage when chief examiners present information needed for grading candidates' work. Detailed information on the relationship between marks

awarded to candidates and the assessment objectives achieved is needed in order to make informed decisions on cut-off scores that reflect standards set in the syllabuses. The system at the Zimbabwe Schools Examinations Council lacks this.

It must be pointed out that examiners need to accumulate sufficient experience to fix standards with reasonable confidence for them to do it right. Chief Examiners and other senior examiners must, after every examination, take note of the standard that the candidates have exhibited in the scripts. They also compare the standards in the sample scripts of the previous examination with current one. Sample scripts are those scripts that represent the cut-off score at each of the key grades of A, C, and E. The Council keeps these scripts for use in future examinations. It is at this stage that this group of examiners can see whether or not the performance of the two groups of candidates is the same or not.

A panel of judges was used in this study to determine the similarity of standards set in question papers from one year to the next. The approach of using experts was used in a number of researches (Elliot & Massey, 1994; Jones & Lotwick, 1979; School Council, 1963; etc). In most of the researches that were looked at, the quality of scripts at a particular grade boundary was scrutinised. For a script that is, say, at grade boundary C, judges would re-mark the script and pass judgement on whether or not it had similar standard of work with grade C script from another Examination board or whether or not the standard of marking was severe or lenient to the work of candidates in a previous examination. None of these researches compared the quality of the question papers, an approach which this study took. In this study I bring in another technique of comparing examination standards through the use of the question papers. I am convinced that analysing the question papers is fundamental because attention is given to the instrument that is used to measure attainment,

therefore, is an effective way of determining standards between subjects and within subjects.

Learning outcomes have been described by many educationists including Kubiszyn and Borich, (1990); and Linn & Grounlund, (1990). Learning outcomes are the attributes of performance and these can also be described as assessment objectives. Learning outcomes are what students are expected to know, understand and manipulate, etc. It is that which a student can demonstrate that he/she has learnt. If the learner can demonstrate that an aspect has been learnt, it is that aspect that can be assessed. In Geography 2248 and Integrated Science 5006, the learning outcomes are found in the syllabuses. It has been argued earlier on that one of the reasons for the localisation of syllabuses from the overseas examination boards was to develop syllabuses that were relevant to Zimbabweans and that had assessment objectives. Syllabuses that have assessment objectives show what students and teachers should focus on in their preparation for examinations. An examination board that has such syllabuses can be described as running transparent system of examining students. The Integrated Science 5006 syllabus states that students must demonstrate: knowledge and understanding of scientific instruments, apparatus, terminology, units, facts and laws, convention symbols, phenomena, definitions, concepts, models, techniques of operation; their ability to extract information, use data to recognise patterns, formulate hypotheses, translate information from one form to another, explain facts, observations, apply scientific principles; and experimental skills through the planning, organising and carrying out of experimental investigations, make accurate, systematic observations and measurements, draw conclusions and make generalisations from experiments.

The Geography 2248 syllabus states that students must demonstrate: knowledge and understanding in the context of scale and areas, processes underlying physical and human landscapes, environmental inter-relationships, definitions; skills and techniques of observation, recording and interpretation, presentation of data, communicating information; some judgement and decision-making in the evolution of patterns in human Geography; and evaluate solutions to environmental and socio-geographic problems.

Whereas in the previous syllabuses described earlier students were not aware of assessment objectives, the Geography 2248 and Integrated Science 5006 syllabuses are elaborate on these. The syllabuses go further to inform stakeholders of the percentage of marks which each assessment objective is given in an examination. There is a relationship between the learning outcomes expected of students who study Geography 2248 and Integrated Science 5006. In the two subjects, students are expected to demonstrate that they have acquired knowledge, that they can comprehend the expected knowledge, apply the learnt knowledge to other situations and to draw conclusions from given situations. This similarity forms the basis of comparing standards set in the two subjects over a period of three years.

Learning outcomes or that which teachers expect students to demonstrate that they have acquired after a learning process were developed by a team of experts led by Bloom and Krathwohl (Gronlund, 1985). They developed a taxonomy of educational objectives. These consist of "a set of general and specific categories that encompass all possible learning outcomes that might be expected from instruction." p.33.

These objectives have become a reference point in educational processes because they provide specific categories of behaviour of students who would have

successfully gone through a learning process. It is emphasized by Gronlund (1985) that each category of learning outcomes includes the behaviour of the lower level. An example is that comprehension includes the behaviour at the knowledge level; the application includes that behaviour at both the knowledge and comprehension levels. These objectives, as can be seen above, have become the basis of the assessment regime that is used by the Zimbabwe School Examinations Council in the subjects of Geography 2248 and Integrated Science 5006.

Kelly (1999) supports the link between specified objectives and the testing of performance of students, noting that the aims and objectives movement that was brought about by those people who advocated for it were concerned with the vague, imprecise purposes that characterised the work of teachers. Some of the proponents of the objectives movement quoted by Kelly (1999) are Bobbitt (1918), Davies (1976), Bloom (1956) and Krathwohl (1964). These proponents of the use of assessment objectives found out that parents and students wanted education in schools to be imparted in a way which they could understand. Kelly (1999) points out that "....the link between the pre-specification of objectives and the testing of performance has been a close one." (p.58).

The closeness mentioned here shows that there needs to be a benchmark against which candidates' performance is measured. That benchmark, in the form of assessment objectives, can remain the same across subjects regardless of the content presented. Mager (1962) points out that a statement of an objective is useful to the extent that it specifies what the learner must be able to do or perform when he/she is demonstrating his/her mastery of the objective.

Kane (1998) puts standards setting methods into two categories. These are test-centred and examinee-centred methods. As the name of the first approach suggests the test-centred methods are those where judgements are made on the test that is administered to candidates whereas the examinee - centred approach relies on judgements that are passed on performance of candidates. It must be borne in mind by the reader that both approaches use judgements, a technique which this study used as well. In both approaches it is clear that before one passes a judgement, a point of reference to indicate what the judge means by a standard is required. A minimum level of achievement must also be known. In other words a performance standard and a cut-off score which reflects each of the performance standards must be known. However, Glass (1978) argued strongly on the point that standards are arrived at in an arbitrary manner. He points out that

> To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. Arbitrariness is no bogeyman, and one ought not to shrink from necessary decisions because they may be arbitrary. However, arbitrary decisions often entail substantial risks of disruption and dislocation. Less arbitrariness is safer. (p.258).

Arbitrariness in determining test scores implies that the scores are arrived at in an illogical, subjective, or uninformed manner. The use of defined criteria such as assessment objectives in setting question papers and in the marking of written work by candidates takes out arbitrariness from the determination of criterion scores. We, therefore, strongly argue that it is not the way used to arrive at the marks and grades awarded to candidates who write the Zimbabwe School Examinations Council

examinations. Moreover, where marking is done after thorough coordination of markers and the known criterion that must be awarded a mark is defined, there would be no arbitrariness. The argument here is that the use of judges in making decisions on cut – off scores does not necessarily imply that the decisions were arbitrary. The use of known criteria when marking scripts and the need for consensus decision on what makes a standard are ways of ensuring that the determination of a standard is not arbitrary. Where a question paper indicates the number of marks that are awarded to each question and where a panel of experts grades candidates using information referred to earlier, it shows clearly that judgements are based on tangible evidence of performance and not arbitrary judgements. The award of marks and consequently grades to candidates who wrote the Geography 2248 and the Integrated Science 5006 was done scientifically. In other words, there is a reason for a mark that was given.

The position taken in this study is contrary to the thinking of Glass (1978). The development of criteria is one sure way of making judgements get focussed on the same predetermined/known performance standards and that can do away with the arbitrariness referred to by Glass (1978).

Kane (1998) proposes that " ... an important first step in designing a standard-setting procedure is for an assessment program to choose between an examinee centred and a test-centred approach." (p.138).

I strongly argue that one does not have to choose one approach because both approaches can be used in order to come up with better judgements. The approaches must be seen as complementary and not exclusive. Judges can review the tasks on a

test which candidates are expected to do and still can go on to evaluate the performance as expected from the set performance standards.

It is important to point out the action verbs that describe learning outcomes because the verbs help readers understand how students acquire the learning outcomes. Kubiszyn and Borich (1990) list the descriptors of the learning outcomes as shown in the table below.

Table 3. Verbs That Describe Learning Outcomes

Level of learning outcomes	Verbs
Knowledge	Define, describe, identify, label, list, match, name, outline, recall, recite, select, state
Comprehension	Convert, defend, distinguish, estimate, explain, extend, generalise, summarise, infer, paraphrase, predict
Application	Compute, demonstrate, develop, employ, modify, organise, operate, prepare, produce, relate, solve, and transfer
Evaluation	Contrast, Conclude, appraise, defend, criticise, justify, support, validate, interpret

These descriptors were used to categorise the questions in the two subjects into the relevant assessment objectives. A panel of experts was used in this study to judge whether or not learning outcomes mentioned here were comparable from one year to the next within a subject and between the two subjects. As has been mentioned

earlier, this was one way of determining the comparability of standards within each of the subjects and between the subjects as well.

It is important to note that any debate on standards, whether in regard to education or in specific subject areas, must be informed by significant, reliable and valid evidence on what pupils know and can do rather than being based on subjective impressions of times gone by. The use of learning outcomes helps in pinning down what it is that is being judged in area of examination standards.

The point made here is the reason why this study based its investigations on the quality of question papers and the grades that were awarded to candidates over the three-year period.

The School Council of the United Kingdom pointed out in its report in 1963 that

> One danger inherent in examinations is that they tend towards conservatism on the part of the examiners. Similar types of questions occur year after year, and often, it must be admitted, because there is an outcry if novel questions are included. (p. 12).

Though this could have been a justified observation in the UK in 1963, the question setters in Zimbabwe are encouraged to set novel questions. The examinations under study had been running for six years when this study was embarked on. A scrutiny of the question papers shows that similar questions do not occur year after year. It is, however, correct to say that the questions from year to year must come from the topics in the syllabus.

As I move into the next area of discussion I need to point out that this section has indicated that standards setting in the area of educational measurement is judgemental. This is a very important point. For the judgements to be sound they have to be based on criteria and norms. I am convinced that the use of criteria is a technique that can render itself to being used by a different group of experts and they would come up with the same decisions. Where criteria are missing then standards could be determined in an arbitrary way.

Interpretation of Test Scores

This study is to an extent concerned with giving criterion-referenced interpretation to the grades obtained by candidates. It is a fact that a bald grade says nothing about a student's strengths and weaknesses. Criterion-referencing provides information about achievements whereas norm-referencing does not tell anyone what it is a candidate knows and is able to do except to state that a candidate has done better or worse than some other candidate. Murphy and Torrance (1988) argue that when criterion-referenced tests are properly designed and conducted, they provide information about what candidates have or have not achieved in a particular field of study. It is important, therefore, to establish by the use of a panel of experts the criteria which the examination question papers in Geography 2248 and Integrated Science 5006 possess and then find out whether or not such criteria can be used to describe what it is candidates who write the examination question papers should be aspiring to achieve in terms of the descriptors of performance.

Hall (1989) points out that "if students' performance against the criteria used for awarding of grades was made explicit, this would be more helpful to students and employers alike." (p. 19.)

This approach to assessment can benefit the students in that while they are at school they will be aware of the areas that they need to improve on. This is the diagnostic use of assessment. This means that areas of strength and weakness are identified and pupils would improve on those areas of weakness and also keep doing the good work. If grades can be described in terms of the skills the candidates possess then it would be much easier to compare standards in different subjects or standards in each subject from one year to the next. When criteria have been identified and test scores are reported to reflect how examinees have achieved those criteria, we have what is called criterion-referenced interpretation of test scores .

Cizek (1993) points out that it was Nedelsky who, in 1954, was the first to talk about absolute standards as opposed to the use of relative standards. Nedelsky cited by Cizek (1993) states that

> The passing score is based on the instructor's judgement of what an adequate achievement on the part of a student and not on the performance by the student relative to his class or to any other group of students. In that sense the standard to be used for determining the passing score is absolute. (p. 97.)

The emphasis here is that the interpretation of a test score in this light is one based on criterion-referencing. Cizek (1993) goes on to point out that the most popular methods

of interpreting test scores is called the compromise methodologies because both normative expectations and absolute judgements are used. This means that these methods are complimentary. It has been mentioned earlier that the Zimbabwe School Examinations Council uses the two approaches but the link between that which the question papers test (assessment objectives) and the grades awarded is not shown by the Examinations Board. As has been pointed out in Chapter I, a bald grade of A, B, or D fails to communicate to the users of examination the competency which a holder of that grade possesses.

Standards can also be seen from the point of view of how an examinee performs in relation to other examinees. In other words, without other examinees the level of achievement of one particular person becomes very difficult to define. The difficulty with this type of standard setting is that once the calibres of the examinees change so will the standard. A pass mark may be pulled down because the candidates failed to achieve the expected marks. In doing so the standards are altered from one year to the next. A standard in such tests known as norm-referenced (Cizek, 1993), is set when a test has already been written.

The stanine system of interpreting scores was developed during the Second World War and has continued to be used (Cohen, Swerdlik, & Phillips 1996). It is a system of standardising scores. The system uses the mean score achieved by a group tested and the standard deviation to put the scores into nine units. Performance can then be reported on a scale of 1 to 9. The system has the advantage that the single digit values

can be manipulated computationally. However, Hoffman (1978) cited in William (1996) points out that "To compress all our information about a single candidate into a single ranking number is clearly absurd-quite ridiculously irrational. And yet it has to be done." (p.290.)

This is viewed as irrational because what a student knows and is able to do cannot be explained clearly to stakeholders by a single digit. The method of reporting performance seem to have been accepted by stakeholders because it has been used for a long time yet quite a lot of information about candidates' performance gets hidden in the digits that represent standards.

It can be argued that the perception of standards in examination boards seems to have somehow been a mixture of these two ways described above. Christie and Forrest (1982) point out that the traditional definition of standards has been the maintenance of a balance between what candidates accomplish by reference to the criteria in a syllabus and by reference to the achievements of other candidates. The view taken in this study is the one expressed by Christie and Forest (1982).

Though norm-referencing results of a test fails to tell us more than the position of one candidate in relation to the others, the technique is useful in indicating to the standard setters how individuals or cohorts performed in relation to others.

Since in criterion-referencing there are targeted criteria, the method provides vital information to the users of examination results. It must, therefore, be possible to

pick on the assessment objectives in Integrated Science 5006 and Geography 2248 examinations and check on the achievements on these objectives by candidates

It should be possible to use these objectives to interpret achievement in an examination in terms of what a candidate knows and is able to do. The idea being presented in this study is to describe grades in terms of what has been achieved by candidates. In doing this the levels of attainment of a candidate who has been awarded a particular grade would be known. This study should describe the grades awarded in terms of standards of performance as indicated in the assessment objectives.

It must be mentioned that criterion-referencing has one major weakness. This is the aggregation of marks as this usually results in the "trade-off" of high marks in one task for low marks in another. What this means is that marks on a test are added up and the total does not indicate the weaknesses of candidates on particular skills or content. The low marks from one particular task are added to high marks from relatively easy tasks, hence the “trade offs”. This point is an argument against the use of criterion referencing in interpreting test scores. I present the argument that assessment objectives must be carefully weighted to reflect their relative importance. Once this has been done the “trade off” problem should be eliminated. The weighting of the assessment objectives must be carefully done so that candidates are not awarded high grades in a subject after they have only demonstrated competency on low order skills such as knowledge. Below are the weightings of the assessment objectives of the two subjects as indicated in the syllabus documents of the years under study.

Table 4. Weighting of the Geography Assessment Objectives

Geography 2248	Paper 1	Paper 2	Syllabus
Skill			
Knowledge with understanding	40%	30%	35%
Application of skills	40%	40%	40%
Judgement and Decision making	20%	30%	25%

Table 5. Weighing of the Integrated Science Assessment Objectives

Integrated Science	Paper 1	Paper 2	Paper 3	Syllabus
Skill				
Knowledge with understanding	70%	70%	0%	
Handling information	30%	30%	0%	
Experimental skills	0%	0%	100%	
Paper weighting	30%	50%	20%	100%

The percentages given in the table relate to marks. The syllabus documents also indicate the paper weightings as shown in the tables above. The problem as can be seen in the Integrated Science table above is that there is no break down of the percentage of skills for the combined papers.

Whenever interpretation of scores is done it is important to know how reliable and valid the tests were in measuring the performance of candidates. Reliability and validity are two important elements of tests that must be of concern to all those

interested in educational measurement. Many authors of educational measurement (Popham 1990; Nikto, 1996; Wood, 1987; etc) have defined validity and reliability. Validity is the degree to which a test assesses what it is intended to assess while reliability is a test's consistency in yielding the same or similar scores when a test is written under the same conditions. Reliability coefficients can be obtained when

(a) candidates complete the same tasks on two different occasions,
(b) candidates complete different but equivalent tasks on the same or different occasions, or
(c) two or more teachers mark candidates' performance on the same tasks.

Reliability is useful when one decides how much confidence to place in the interpretation of assessment results.

The reliability of examinations must be discussed in the light of the marking procedures. The way scripts are marked have a bearing on the standards of the examination. The two subjects under study were tested using two techniques. These are the multiple choice and the essay techniques. The answer scripts for the multiple choice question papers are electronically marked. Unlike the essay type answers that are marked by humans, this paper does not have a problem of inter-marker reliability. Certain steps are taken to improve the inter-marker reliability in the essay type papers. These are described below.

At the Zimbabwe School Examinations Council the reliability of marking essays is improved by the use of detailed marking schemes which leave the markers in no doubt of the answers expected from the candidates. The construction of marking schemes of the two subjects under study is not the responsibility of one individual. The marking schemes are constructed during the time of test construction by a panel

of experts in the subject. This ensures that all the possible answers are included in the marking schemes. Further to this, co-ordination of marking meetings are held before the start of the marking. The purpose of these meetings is to ensure that

a) all the possible answers have been included in the marking scheme;
b) there is agreement between markers on where marks will be awarded; and
c) there is consistency in the marking of the scripts.

On the first day of the coordination meeting the leadership of the marking exercise discusses the marking scheme. Care is taken during these discussions not to transfer marks from one part of a question to the other as this could affect the weighting of the assessment objectives that was agreed upon at the setting of the question papers. A team leader checks the work of each marker and the Assistant National Chief Examiner in turn checks the quality of marking of the Team Leader. The Chief Examiner checks the work of the Assistant National Chief Examiner. This checking mechanism is thorough and is there to ensure that there is a high inter-marker reliability.

The reliability of tests refers to the consistency with which a test measures what it is supposed to measure. A ruler, for example is a consistent instrument used to measure lengths. It is a reliable instrument in what it measures. Two metres of a piece of wood will still be two metres the next time you measure it as long as no one has cut a piece off. However, examination question papers may not be as reliable as a ruler. This emphasizes the difference between reliability in the physical sciences and in educational measurement. Such differences prompted the researcher to look at the question papers in Geography 2248 and Integrated Science 5006 to determine whether or not they were consistent instruments to measure candidates' performance.

Nhandara (1994) carried out a study to determine the reliability and validity of a fifty-item multiple-choice Geography 2248 test. A sample of 809 students from the Harare educational region in Zimbabwe was used. A 15-member panel validated relevancy and representiveness of the items. The internal consistency of the test was 0.79. This was a high internal consistency of the test. The discrimination indices were above 0.20. The discrimination indices that are above 0.20 indicate that questions in a test were discriminating well, poor candidates from the able ones. The panel of judges generally agreed with the content classification.

As mentioned earlier, validity is the extent to which an examination or a test does what it is designed to do. There are different kinds of validity because validity depends on the purpose of a test (Weirsma & Jurs, 1990, Cohen & Manion, 1992), but the concern of this study was content validity. In other words, experts checked whether or not the domains tested were the ones in the syllabuses. Some of the types of validity are construct, face, concurrent, and criterion-related. The definition of validity refers to the outcome of a learning process. It refers to the syllabus domains, which refers to the content that is specified in the syllabus and the skills that must be acquired to demonstrate competency at doing things. Validity, therefore, must give the assessor evidence that supports the meaning of test scores. Some test instruments may be worded in such a way that they are only accessible to very few candidates who have a very high vocabulary. In that case a test may not be testing what it purports to test. The two fundamental requirements of any assessment are its validity and its reliability (Linn & Gronlund, 1995). The two have an important bearing on the soundness of an assessment that is used. It is important to understand validity as a process in that it is concerned with the accuracy of measurement whereas reliability is concerned with the precision of measurement. It is quite possible to measure something with great precision but not be an accurate measurement of what is

intended to be measured. Validity is also concerned with the appropriateness, meaningfulness, and usefulness of inferences that are made in interpreting test scores. Reliability uses quantitative techniques for analysing the assessment whereas validity uses the qualitative analysis. The objective of this study was to establish whether or not the measuring instruments in Geography 2248 and Integrated Science 5006 had the same content validity over the three-year period under study. It must always be remembered that content validity places emphasis on the relationship of a test to the syllabus learning domains. The content validity of examinations can be determined by answering quite a number of questions. During the setting of question papers the setters must answer the following questions: Are all the question setters absolutely clear about the syllabus content domains assessed in the examinations?; Are the setters clear about the assessment objectives which the examination should assess?; Do the setters, themselves, understand the meaning of syllabus objectives?; and Do all the setters understand the concept of validity in assessment?

One can see that these questions focus on the examination setter's understanding of the relationship between the syllabus content and the examination. These are the important points to always remember when one is concerned with the comparability of examination standards.

A research paper that was produced by the Senior Secondary Assessment Board of South Australia (SSABSA, 1998) pointed out that questions may not always test one assessment objective. The research found out that a question could have in it two or more assessment objectives. An example was given of where a question assessed a

candidate's knowledge of Biology as well as their problem solving skills. This, it must be emphasised, is normal in the assessment world because questions can be set to evaluate how a candidate can develop an answer from a simple construct to a complicated one but what is important is how the marks for that question are balanced in terms of the low and high order skills. Failure to make the appropriate balance can lead to skewed standards.

The discussion on reliability and validity shows that the two are very important when one focuses on standards in examinations. If the testing instruments are not valid and the marking of scripts not reliable then the examination as a whole would not represent any standard at all.

Comparability of Standards

There are various methods which are used to monitor standards in public examinations (Bardell, Forrest, & Shoesmith, 1978; Christie, & Forrest, 1981; Forrest, & Vickerman, 1982; Matthews, 1985). Some of the methods are as follows: subject pairs analysis, comparison of grades obtained by examinees from one year to the next, and the use of reference tests to establish the reliability of the grades awarded in an examination. One fundamental feature of all these methods is that they analyse the outcome of examinees' work. In other words, they compare how tests successfully "sort out" examinees into different grades. Christie and Forrest (1980) warn that any conclusions drawn about comparability of standards must depend upon prior establishment of the extent to which it is the same attribute which is being evaluated.

It needs to be emphasized that the subjects under study have the same attributes that are the basis of the comparison. These are assessment objectives. These have been discussed earlier.

Kolen (1999) investigated the comparability of test scores on two different types of tests. The pencil and paper test and the computerised mode of a test. The investigation was centred on the question: how can a testing organisation ensure that the scores earned on the pencil and paper test indicate the same level of achievement as scores on the computerised test? Though the article is not directed to comparability of standards in secondary school education there are similarities in what the article dealt with to what is investigated by this study. The framework given by Kolen (1999) helps to identify threats to comparability in assessment. The areas where there are threats to comparability can come from the differences in test questions, test scoring, testing conditions, and in examinee groups. Kolen (1999) concedes that test specifications are vitally important in order to arrest the problem of differences in test questions. The Geography 2248 and Integrated Science 5006 syllabuses indicate specifications for the tests. The scoring of the tests in the two subjects was the same within each of the subjects over the three-year period. Moreover, the two syllabuses are graded on the same General Certificate of Education Ordinary Level standard of grades A to U. The testing conditions were the same for the candidates who wrote the examinations because they were administered according to the same laid down regulations.

A considerable amount of literature was produced through the Joint Matriculation Board (JMB) in the UK on the subject of comparability of examination standards ((Bardell, Forrest, & Shoesmith, 1978; Christie, & Forrest, 1981; Forrest, & Vickerman, 1982;). It must be borne in mind that at one time there were more than nine General Certificate of Education (GCE) Examination Boards in the UK (now the Examination Boards which offer the General Certificate of Secondary Education (GCSE) are in regional groupings). There was, therefore, great concern over whether standards were being maintained across subjects offered by the different boards.

Bardell, Forrest and Shoesmith, (1978) compared the percentages of examinees who were awarded particular GCE grades by different examination boards. Though Bardell, Forrest and Shoesmith, (1978) came to the conclusion that two or more boards differed in their standards it was impossible for them to say which of the standards was the correct one. This is because the methods of monitoring standards used did not have concrete external criteria which can become a point of reference for all the boards.

Bardell, Forrest and Shoesmith, (1978) also discussed the technique of using monitoring or reference tests when investigating the maintenance of standards. Regression analysis was used to relate performance on the actual examination to the performance on a reference test. Though the technique can give some indication of a relationship between reference test grade and a grade obtained in an actual examination the reference test need to be a reliable and valid measuring instrument. Items which constitute a reference test must be obtained through techniques which provide sample

free statistics. There could also be a problem with examination boards which offer different syllabuses when this method is used. This is a stumbling block to those who advocate for the use of a common paper in an examination in order to determine the level of competence of examinees across all boards.

A qualitative approach that also has some quantitative aspects is cross moderation. This technique is where experts, in this case examiners, determine whether grades awarded by one board are comparable to levels of attainment set by another examination board. This technique is problematic although it enables the decisions taken by examiners to be substantiated by the quality of scripts. Examiners from Board A could say that Board B set a standard which was too high while Board B says examiners from Board A set standards which are too low. This happens whenever there is no external criterion that must be a point of reference when setting standards. Yet another technique is the use of examiners to check whether standards in each subject are being maintained from one year to the next. In such a case examiners are asked to remark the previous year's scripts and compare the standard to those of the current year.

Bardell, Forrest & Shoesmith, (1978) hold the view that the use of chief examiners in standards setting is the most fruitful and sensitive one.

However, Matthews (1985) who criticises examination boards' reliance on the skills of chief examiners in making consistent judgements on the quality of work at cut off scores overlooks the fact that setting of educational standards is a judgmental issue.

Errors do exist in a process of that nature but judges must strive to reduce those errors. The pivotal role of Chief Examiners in carrying standards from one year to the next has already been pointed out above. The inconsistency in making their judgements can be reduced by providing them with as much information as possible about test items.

Forrest and Vickerman (1982) carried out a subject pairs analysis of the standards set at the GCE level between 1972 and 1980. They compared mean grades of subjects and found out that the relationship between many subjects at Ordinary level between these years was remarkably stable. An assumption that is made when using this technique is that the shapes of the distribution of grades in a pair of subjects are similar. In criterion-referenced tests it is not the distribution of the grades which is important but rather the ability of an examinee to achieve laid down criteria.

If one carried out a subject pairs analysis over a period of eight years as was done by Forrest and Vickerman, (1982), one needs to ensure that there were no syllabus changes during that period. If the syllabus changes then the comparison of the subject grades will not be based on the same content and criteria.

Wood and Skurnik (1969) are precise on the problems of maintaining standards in public examinations. They point out that examinations vary in quality because examination boards do not have precise indicators of the quality of test items which constitute the examination papers. Instead they use subjective assessment of the quality of work and find it very difficult to verify their impressions. Item Response Theory can

provide objective measuring instruments. This technique of maintaining standards provides judges with information about an item and the ability which an item can measure.

The use of the Item Response Theory technique of maintaining standards in educational measurement is important in the area of detecting item bias, providing item statistics such as item difficulty, item discrimination, and guessing. These statistics are not sample dependent unlike in the classical test theory.

Van der Linden (1981, 1982) has argued strongly for the use of the item response theory to solve the problems experienced in the area of educational measurement. The argument presented is that the interpretations of how an item functions are usually misconstrued because of lack of information about that particular item. Van der Linden (1982) compared the standards set by judges using the Angoff and the Nedelsky methods to those where Item Response Theory was used. He found out that the decisions of borderline scores reached at after using the techniques of Angoff and Nedelsky were not compatible with the probability of success identified by the Item Response Theory. The reason for this difference in the standards set is that the judges who use the Angoff and Nedelsky techniques must have in mind the qualities of a borderline examinee. The decision of a borderline examinee is based on their own conception of the properties of items and this conception may not be the way the items actually functioned.

The method has the weakness that in longer tests the differences in the decisions made tend to average out. The end result is that low differences can be observed yet in fact there could be larger differences in decisions reached at each item. However, the method is powerful in the area of standards setting.

Van der Linden (1981) compared the use of Item Response Theory with the Cox and Vargas validity index (Dpp) of criterion-referenced tests. He/she criticises the work of Cox and Vargas because the discrimination index is population dependent just as the classical theory discrimination index. Van der Linden obtained the Cox and Vargas Discrimination Index through the use of the three-parameter logistic model. In this case the index consists of an Item Characteristic Curve and the difference between the pretest and post-test mastery distribution. The curve is independent of any distributional characteristics of the population scores but only reflects the properties of the item.

The work of Cox and Vargas has the weakness that it mixes two sources of information. These are the characteristics of the item and the differences between the pretest and post-test mastery distributions and then blame the former on the peculiarities of the latter. Item Response Theory has the advantage of not mixing these two issues because it gives item statistics which are not dependent on the population used. Van der Linden observes that the virtue of this application of item information functions also lies in the fact that it provides population - invariant test design based solely on the characteristics of the items.

Seddon (1987) points out that tests that are constructed after their parameters are known are more effective than those chosen at random from a domain. Identification of items whose parameters are known can make it easier for standards setters to maintain the standards they set.

Item Response Theory has been used by Methuen, Chih-Fen, and Burstein, (1991) to detect item bias. It must be remembered that item bias affects the validity of a test and validity of tests affects standards that are set from one year to the next. The assumption of the model used is that there is invariance of measurement parameters for different subgroups of the population tested. When a comparison of the curves, which describe the probability of a correct answer for a given ability across groups, is done and if a large area between the two curves exists then an item is biased. If particular standards in a test to be constructed are to be maintained then it is important that biased items can be discarded at that stage of test construction. However, Methuen, Chih-Fen and Burstein, (1991) point out that although the Item Response Theory assists in item bias detection, it is difficult to know the kind of bias an item has. Experts would be needed to identify the reasons for the bias.

One problem in the maintenance of standards through the use of the Item Response Theory was identified by Goldstein (1983). He found out that the use of the Rasch Model to measure the properties of items is weak because test properties change over time. He compared two tests which were administered to eleven and fifteen year olds

between 1948 and 1970 and also between 1955 and 1970. Goldstein identifies duality between attributing change in an item parameter value to change in the population response and in the characteristics of an item. This is particularly so in a reading test where one may have words which at some time were more common in daily usage than others. Agreeing with Goldstein that there can be item parameter drift, Block (1978) identified the reasons for such drifts as educational, technological or cultural changes. The authors suggest that item pools can be updated by items which are on an Item Response Theory scale in order to keep them abreast with changes in education. This helps to maintain the standards of items in a bank.

Sykes and Fitzpatrick (1992) carried out an investigation into how the Item Response Theory difficulty value can remain stable over consecutive administration of an examination. This research was carried out on the background of having difficulty estimates which varied because of the position of test items in a test. The authors cite Yen (1980); and Eignor & Cook (1983) who found out that some reading comprehension items became more difficult when put later in a test. They found out that there was no relationship between changes in difficulty values and the position of test items in a test. They attributed the shift of the difficulty values to shifts in curriculum emphasis and not in the position of items. This problem will not be anticipated in this study since there was no curriculum shift at all.

The comparability of standards is whereby performances achieved by candidates in one subject in one year are compared to the next. The comparison can also be between two subjects. Achieving comparable standards, therefore, refers to the

award of similar grades to a candidate who exhibits the same level of mastery of identifiable skills across subjects and the maintenance of the same quality of questions within a subject from one year to the next. Comparability of standards also refers to the number of candidates achieving particular grades within a subject from year to year. Examination Boards use statements of attainment or grades to communicate students' achievements in examinations to the stakeholders. The Zimbabwe School Examinations Council uses grades. The grades that are used at the General Certificate of Education Ordinary level are A,B,C,D,E and U. The highest grade that can be achieved is A while U is the lowest. D,E and U are fail grades.

There are many reasons for carrying out comparability studies. It is important to take note of some of the reasons here. Comparability studies alert us to potential problem areas that merit monitoring and further investigations. For example, Willmott (1980) revealed that Technical Drawing examination results were much worse than the results in other subjects for the GCE examination in England and Wales. In such a case, more investigations would be needed to determine why the results were much worse. Willmott's research did not. The School Council Bulletin (1963) suggested that the direct comparison of the performance of candidates be used to check on standards.

Comparability studies also help in detecting shifts in trends. These studies further provide insights into what is happening to the curriculum and the means by which it is assessed.

Adams and Phillips, (1988) conducted an inter-group comparability study for the Joint Council for the General Certificate of Secondary Education (GCSE). The objective of the study was to compare the grades awarded in the subject of

mathematics by six examining groups in England, Wales and Northern Ireland. Examiners were called in to pass some judgements on the grades awarded at the boundaries A/B, C/D and F/G. It is worthwhile to note that the comparisons in the grades were made against the background of differences in the detail of the assessment schemes of the different boards. Each of the examining groups presented sixteen scripts at each of the boundaries. The examiners were asked to state whether or not a script was a borderline. They also needed to state the grading standard they would have agreed upon. The results of that research showed that the Northern Examining Group had stricter standards while those for the London and East Anglia Group (LEAG) were more lenient. It is important to note that the study also revealed that there

> was the difficulty experienced by scrutineers in separating the different factors which combine together to produce the composite concept 'standard'. Their principal concerns were the demands made by the question papers in relation to the level of achievement expected at the different grade boundaries and the balance of skills and abilities assessed within the papers. (p. 43).

This is a very important observation but in this study on the comparability of standards set at the GCE "O" level Geography 2248 and Integrated Science 5006 the scrutineers focussed on the quality of questions from year to year. They established whether or not the questions reflected the demands of the specification grids in the syllabuses.

Patrick (1996) argues that there is more sense in carrying out comparability studies over a short period of time than over a period of say twenty years. She/he compared

the 1906 and 1951 English composition question papers and found out that the context within which the examinations were set, taken, marked, and graded was different. Changes such as social, demographic, cultural and technological affect the learning environment and examinations also change accordingly. It has been shown above that Goldstein (1983) and Block (1988) concurred with this point. Twenty years ago, Patrick (1996) points out, syllabuses were simply lists of context headings but now they include assessment objectives and criteria. In this study such a difference between the UCLES and the ZIMSEC syllabuses has been noted. This study focused on comparability of standards over a period of three years. According to Patrick (1996) the time period of this research falls within the acceptable limit.

Vengesayi (1991) compared the 1989 questions at pretest stage and the final Grade Seven General Paper examination of 1990. The Grade Seven examination is written after seven years of primary school education. Though this examination is different from Ordinary level examination what is of significance is the fact that Vengesayi (1991) was comparing the standard of questions at the pretest stage and those at the actual examination stage. It was a study that determined the standard of questions. The study revealed that items that were identified as performing well during the pretesting stage did not perform well in the final examination. This means that the examination standard set at the pretesting stage was not the same as that of the final examination. The comparison of standards done by Vengesayi (1991) was based on the syllabus and item statistics. The researcher found out that there were significant differences between pretest results and the final results. This indicated different quality of items used in the tests. He revealed that some items that were in the final 1990 paper had not been pretested in 1989. The correlation coefficient for the discrimination index and the difficulty index for the 1989-pretest paper and the 1990 question paper were 0.162 and 0,0090 respectively. This was a very low correlation

coefficient between the two papers and Vengesayi (1991) accepts the fact that the low correlation could have been a result of the different samples used.

Another technique for comparing standards is known as follow-up studies. This is where a check is made on what students who have gone through the examination are doing, that is types of employment or further education. The study will have to establish whether or not the students whose performance was the same at school will be doing similar work or studies. The weakness of this technique of comparing standards is obvious. Where there are high levels of unemployment in a country such as Zimbabwe, students would end up in jobs that do not reflect their performance at school. Furthermore, at institutions of further education in Zimbabwe where there is usually a scramble for places, many students may end up taking subjects that they were not very good at but rather get into courses where places were available.

Standards can also be maintained by statistical methods. When the quality of students and that of the teaching is unchanged for a period of years and if the number of candidates is reasonably large, the proportion reaching any given level may be expected to be stable. However, the qualitative judgements must not be ignored because the statistical controls must always reflect these judgements. An example of qualitative judgements which Chief Examiners make are that they recommend a particular mark as a cut-off for a grade because of the quality of answers which they would have seen during the marking sessions. These are the points which should not be ignored when standards are set by examination boards.

Comparability of standards as a technique of monitoring standards within and between subjects and between examination boards has been used for a long time

(Bardell, Forrest & Shoesmith, 1978; Jones & Lotwick, 1979; Forrest & Vickerman, 1982; Vengesayi, 1991, Eliot & Dexter, 1995; Massey, 1997; etc). Forrest and Vickerman, 1982 used two methods to investigate subject comparability. These methods are

a) the use of an external test as a reference point in judging the standard of performance on other tests and;
b) the use of the GCE "O" level examination results only.

Nuttall, Backhouse and Willmott cited in Forrest and Vickerman (1982) carried out a comprehensive report on subject comparability using these two methods. They found out that the second method provided a high degree of consistency in the results of comparability. This second method will be used in this investigation. The School Council (1963) also discusses the techniques of establishing standards. They point out that where syllabuses are different it is possible to look for levels of argument, comprehension and style that can be discerned as common standards irrespective of content.

Research carried out by the Joint Matriculation Board (JMB) in 1971 (Forrest & Vickerman 1982) compared the average grades of the candidates who wrote English Language Paper B, Geography and Chemistry at the "O" level examination. The research revealed the same results as those which had been found by Forrest and Shoesmith (1985) when they used a reference test. This showed the reliability of the two methods of monitoring subject comparability.

Research into comparability of standards in Physics and Mathematics carried out by the Northern Examining Authority (NEA) Research Advisory sub-committee (1990) using a sample of 6 500 candidates, revealed that the same standards were not

applied in the grading of the candidates. This sub-committee found out that it was easier for a candidate with a grade C in Physics to get grade C or better in Mathematics. However, the research did not go further to investigate the criteria used in determining the grade cut-off points.

Elliott and Massey (1994) investigated the standards that were set in UCLES International General Certificate of Secondary Education (IGCSE) Foreign Language French (June 1994) and Midland Examining Group (MEG) General Certificate of Secondary Education (GCSE) French (June 1994). The reason for carrying out the investigation was that IGCSE centres were disappointed by the grades that their abler candidates were getting. The method of comparing performance of candidates was cross moderation. The researchers accepted that comparability is a difficult exercise because the demands of the papers and marking schemes that were compared were different.

Five scripts per borderline of A/B, C/D, and E/F were selected and used in the study. Six judges, three from IGCSE and three from GCSE, were used to make judgements on the standards presented. The experts were to make judgements of whether scripts reflected a:

a) typical borderline
b) performance better than the borderline
c) performance worse than the borderline

These three judgements were converted to a numerical scale as 0, +1, or -1 respectively. The levels of agreement between the judges of the boards were checked by the use of a coefficient of concordance. They found out that the overall rater concordance was very high at the E borderline but at the A/B borderline the GCSE judges did not quite agree and so did the IGCSE raters at the C/D borderline. The study showed that there was a disparity in the standards at the GCSE and

IGCSE. The researchers recommended that the IGCSE grading standard be adjusted by as much as a grade in order to bring them into line with the GCSE.

The Standing Research Advisory Committee of the GCE Examining Boards in the UK carried out an investigation into the standards in Advanced Level Mathematics. One of the reasons for carrying out this research was to relate the results obtained on double mathematics to those obtained by the same candidates on the other "A" level subjects which they had written. The comparison of performance was done through the use of mean grades obtained in Mathematics, Geography, Chemistry, Biology, Economics, Computer Studies and Physics. The comparisons were done for candidates in each of the following boards in the UK: The Associated Examining Board, UCLES, Joint Matriculation Board, University of London School Examinations, Northern Ireland Schools Examinations Council, University of Oxford Delegacy of Local Examinations, Oxford and Cambridge Schools Examination Board, Southern Universities Joint Board, and the Welsh Joint Education Committee.

The research revealed that when double mathematics is Pure Mathematics and Applied Mathematics or Pure Mathematics and Statistics the mean grades for those Mathematics subjects were higher than those obtained in Physics, Chemistry, Economics, Computer Studies, Biology, and Geography when these are third taken subjects. The mean grades obtained in their mathematics subject by candidates taking only Chemistry with double mathematics were usually higher than when only Physics was taken.

Jones and Lotwick (1979) investigated the relative grading standards in the 1978 Biology "O" level examinations in the UK. They compared standards across nine examination boards. Experts in Biology made judgements on scripts on the

borderline of grades C and D to determine whether or not the nine boards used the same standard. The experts were asked to indicate on a 5-point scale whether the standard was correct, lenient or severe. It must be pointed out that the nine Biology examinations, which were compared, varied quite significantly. Some boards had only one 2½ - hour paper, others had only one 2-hour written paper, and yet other boards had multiple choice and a practical test. The number of questions in the papers differed from board to board and so did the mark allocation. Nine Chief Examiners from the nine boards were asked to assess the presented scripts. Four independent assessors were also asked to do the same job. Spearman's correlation coefficient of 0.93 was achieved between the independent assessors and the Chief Examiners. Kendall's coefficient of concordance was used to check on the significance of the agreement among chief examiners in judging the grading standards. They found out that the agreement within Chief Examiners and within assessors was higher than it would have been by chance at 1% level of significance. This research shows that comparisons between examination papers of different duration can still be made in order to investigate the level of difference or similarity in standards used. However, this method of checking standards has a disadvantage. Each chief examiner usually regards the standards of other boards as lower than the board she/he represents. Jones and Lotwick (1979) argue that such a bias is normal in any cross moderation exercise but the overall results are valid as long as the exercise is anchored on known criteria.

The common ground of comparing the standards in Geography 2248 and Integrated Science 5006 examinations was the same assessment objectives upon which the tests were based. The view that candidates may obtain the same grades when they have demonstrated different achievements by answering different questions in a test

(School Council 1979) can only be put forward where the questions in a test do not have parallel skills.

It can be argued that it is not possible for a candidate to be awarded the same grades in all the subjects that he/she attempts at a particular level. Indeed, there are a number of factors that could make a candidate get different grades in different subjects. However, Forrest and Vickerman (1982) and the School Council (1979) found out that if a large group of candidates representative of the population wrote examinations in related subjects their mean grades will be the same. This was the basis for investigating comparability of examination standards in this study. It was also the reason why this study on comparing examination standards in GCE "O" level Geography 2248 and Integrated Science 5006 was carried out.

One technique for establishing whether standards are comparable or not is the use of sample scripts. Samples of scripts at particular grades are archived and used as benchmarks in the following examination when fixing standards. The School Council (UK) (1963) supports this technique. They point out that "one way to minimise the possibility of varying standards is to adopt a matching technique in which a number of key scripts are selected as fixed reference points, all other scripts are then matched against the chosen few." p16.

Elliott and Dexter, (1996) produced a report addressed to the people who have the responsibility of maintaining standards for the GCE "A" level examinations. The objective of pursuing the study was two fold. First, it was to establish whether or not there was comparability of grading standards between the University of Cambridge Local Examinations Syndicate and other boards. Second, the researchers wanted to check whether the grading standards in various subjects offered by the University of

Cambridge Local Examinations Syndicate broadly agreed. The researchers used the subject pairs comparison data that the Syndicate routinely produced as well as the other boards' statistics. The authors compared the mean grades achieved in the different subjects. The alpha grades were transformed to numbers as follows:

Upper grade A = 1

Lower grade A = 2

Grade B = 3

Grade C = 4

Grade D = 5

Grade E = 6

Grade N = 7

The researchers used graphs in order to provide visual inspection of the relationship between subjects offered by different boards. The researchers found out that UCLES pass rates A to E were in line with the consensus across the boards. However, standards in General Studies were severe in 1994 at grades A and B but in 1995 they were back in line with the other boards.

When Massey and Dexter, (1995) used coursework element to monitor grading standards in four large entry Midland Examining Group GCSE Science syllabuses, they found out that the correlation with the achievement on the written paper was weak. They blamed the variations between teachers' emphasis in coursework and written papers. That research showed that school effects on results could not be ignored when looking at the comparability of standards. This is true because what was being compared was coursework and public examinations and so school effects become obvious in those circumstances. School effects will not influence the

comparability of standards in Geography 2248 and Integrated Science 5006 because the sample was carefully selected using the stratified random technique.

This chapter has given the conceptual framework to the study in that standards have been shown to be important for what they are, as well as for what they do. In support of this I have explained the term standards and on what they are based. The assessment objectives in the syllabuses of Geography 2248 and Integrated Science 5006 were discussed in this chapter and many authors were referred to in order to illustrate the importance of standards in education systems. In as far as what standards do it is important to look at how the study deals with the interpretation of test scores because the same standards could mean different things to different people. This study now moves to the next chapters using assessment objectives as the bases of determining the existence or non-existence of standards in question papers in Geography 2248 and Integrated Science 5006 over the three year period, 1996, 1997 and 1998.

The discussion in this chapter has also given direction to this study to use experts in the field of educational measurement in passing judgement on the quality of question papers. These experts were qualified teachers who taught the subject at "O" level. They were further trained by ZIMSEC in marking, setting and grading of candidates' work. By the time this research was done, these people had gained considerable experience in examinations. The question papers are fundamental to the study because they are the measuring instruments for standards in an examination. The use of any other group such as parents, students and teachers would not have been appropriate because we would get impressions on standards rather than professional judgements from people trained in the marking and setting of examination question papers.

Furthermore, a look at the methods of interpreting test scores leaves us no doubt that a criterion-referenced interpretation of test scores links what is in a testing instrument, the question paper, and an achieved score. This study compares standards as represented by assessment objectives. It has, however, been established that many authors view criterion-referencing and norm-referencing of test scores as complementary. The methods of grading at ZIMSEC that brought about the grades that were used in this study were based on the two methods.

It is important to point out that the technique of comparing standards known as subject-pairs analysis used by Forest and Vickerman (1982) became the basis of comparing mean grades of the subjects in this study. This technique was the second method of comparing standards after the comparison of skills that were in the question papers.

This study takes into consideration the threats to comparability that were discussed above. It is important to reiterate that the regulations that governed the testing conditions of the "O" level examinations and the test scoring did not change over the three-year period and so these were not threats to this comparability.

It is with this understanding of standards, interpretation of test scores and the techniques of comparing standards that I move on to the chapters that follow.

Summary

The reviewed literature showed that a standard is a performance indicator and can be test centred as well as examinee centred. The standards of examinations that were

compared in this study were attributes of performance upon which judgements could be made by anyone who cared to. The attributes of performance in Geography 2248 and Integrated Science 5006 are anchored on the assessment objectives that are in the syllabuses. The reviewed literature showed that judgement on standards referred to whether or not the standards were above or below the borderline of a grade. This study went further than what has been reviewed in the literature in that it sought evidence on the degree of similarity of the content, assessment objectives, and the relationship of grades that were given to candidates who wrote the Geography 2248 and Integrated Science 5006 examinations in 1996, 1997 and 1998.

The development of the Zimbabwean examinations system involved the training of markers and setters by both the UCLES consultants and local specialists, the setting of question papers, the marking of scripts, the processing of results and the establishment of an autonomous examinations board called the Zimbabwe School Examinations Council. The value that the Ministry of Education placed on examination standards was seen through the route that was taken when localising the examination system.

The reviewed literature did not establish what it was the grades that were compared meant in terms of what candidates knew and were able to do. Some judges who were used in the researches that were discussed above made judgements on whether or not different boards used the same standards. Other than saying that the standards used

were above or below the borderline, the judges did not state what constituted a borderline.

Though there was so much literature on the comparability of standards in the UK, none was found on the comparability of ZIMSEC standards in Geography 2248 and Integrated Science 5006. This study represents the first attempt in the history of the Zimbabwe School Examinations Council to compare standards within Geography 2248 and Integrated Science 5006 as well as between the two subjects.

It is possible to use public examination results in order to assess comparability of examinations standards. Public examination results have been found to yield more reliable results than the use of an external reference test.

Comparability of standards between subjects is of paramount importance to the credibility of examinations. The fact that examination boards such as the JMB, UCLES, Oxford and Cambridge continue to work on comparability of examination standards shows the importance of this area to examination boards.

CHAPTER III: RESEARCH METHODOLOGY

Introduction

This chapter looks at the composition of the population of candidates who wrote the Zimbabwe General Certificate of Education Ordinary level examination in Geography 2248 and Integrated Science 5006 over the three-year period under study. The sample that was taken from the population is also described in this chapter. The correlational and the survey research designs, and the instruments that were used in this study were also discussed.

Population

Many candidates at the "O" level opt to write the Geography 2248 and Integrated Science 5006 examinations. Table 6. below shows the candidate entry in the two subjects over a period of three years.

Table 6. Candidate Entries Over The Three-Year period

Year	Geography 2248	Integrated Science 5006
1996	123 082	146 461
1997	138 102	146 363
1998	151 290	164 843

Out of the entry figures shown in Table 6 there were about 110 000 candidates who wrote both Geography 2248 and Integrated Science 5006 "O" level examinations in the

three years, 1996, 1997 and 1998. These candidates came from different types of schools in Zimbabwe. The differences in the types of schools in Zimbabwe emanate from differences in geographical locations and responsible authorities. There are government, private, church, council, farm and mine schools. This is a Ministry of Education classification. All the secondary schools in the country were keyed into the Microsoft Excel computer programme. This is a spreadsheet. The spreadsheet made it possible to have the schools' examination numbers, the name of each school, the region in which each school is located, the type of each school and the number of candidates at each school that wrote Geography 2248 and Integrated Science 5006. All the columns on the spreadsheet except that of type were filled in when the data entry sheets were sent out to the Examinations Officers in the nine Ministry of Education regional offices. The nine Officers were to confirm the classification of schools. A copy of the list that was sent to the regional offices for the confirmation is shown as Appendix F.

The locations of the schools differ as some are in the rural areas while others are in urban centres. Government schools in urban centres are further split into Former group A and Former group B. The former group A schools are found in the low population density areas while the former group B are in high population density areas. This classification is also linked to the colonial history of the country. Schools were split along racial lines before the independence of Zimbabwe in 1980. The group A schools were for children of European descend (whites) while the group B were for the African children (blacks). The categories are still to date referred to as former group A and former group B. The private school category is where the high fee paying schools were

found. Church schools have church organisations as responsible authorities. The main characteristics of council schools in the urban centres are the same as the government former group B schools while those in the rural areas do not differ from the government schools there. Accordingly, in this study the council schools, though they are private institution, have been classified as government former group B urban or rural schools depending on their geographical location. Farm and mine schools were, in this study, put in the same category because their main characteristics are the same. Below is the coding system that was used in this study.

Table 7. School Categories

CATEGORY	CODE
Government School Former Group A	A
Government School Former Group B	B
Church	C
Independent High Fee Paying Schools	D
Farm & Mine Schools	M
Government & Council Rural Schools	R

The schools that offered the two subjects to candidates in the period under study are listed in Appendix F, which is a sample of the forms that were sent to all the nine regions. The appendix also shows the educational regions a school was found, the category that a school lies in, the school population, the geographical location and the

number of candidates that entered the two subjects for the examination. It can be noticed from the data given in Appendix F that the schools that offer these two subjects have different population sizes. The candidates who wrote the two subjects were identified using the Graded Candidate lists that were produced by the Zimbabwe School Examinations Council. A copy of this document was attached as Appendix H. Since at some centres some candidates could only write one of these subjects, the sample was made up of only candidates who wrote the two subjects.

It was important to describe the characteristics of each category of school in the population. The categories of the schools in Zimbabwe differ in the physical, financial and human resources that are available to them. Government Former group A schools had more classrooms and laboratories than Former group B and rural schools. It was quite common during the period this study was undertaken to have students in rural schools and government former group B schools to experience shortages of textbooks, furniture and qualified teachers. Such problems obviously have an impact on the performance of candidates. The School Development Associations (SDAs) in private and former Group A schools had more money than those in former group B and rural schools. The SDAs are associations of parents which have the authority to fix a development levy that each parent is obliged to pay. The associations at former group A schools generally charged high levies than former group B and rural schools. This was because the urban parent has better financial resources than the rural parent. The former group A associations had the financial resources to employ teachers to add to the complement provided by the Ministry of Education. Private schools charged very high

fees and had the financial resources to provide qualified teachers as well as buildings for the needs of the school. It is in rural schools that one found untrained teachers and inadequate classrooms and laboratories. Even if laboratories were available at rural schools getting the equipment and the chemicals at some schools could have been a problem because of limited financial resources. Qualified teachers do not like to teach at schools in the rural areas because in general the schools do not have running water, electricity and descent houses.

The Sample

It has been shown above that the population of the candidates who wrote the Geography 2248 and Integrated Science 5006 had clear strata. Leedy, (1980) points out that proportional sampling design must be used when the population is not homogeneous but "..is composed of layers (strata) of discretely different types of individual units." (p. 119).

This is the sampling design that was used in this study.

The population was put into the classes mentioned above and a sample of two percent from each strata was randomly selected. Orlich, (1978) recommends that for a population that is over 75 000 one must use a sample of at least 383. This represents 0,51%. However, Cohen and Manion (1992) and Erickson and Nosanchuk (1988) point out that there is no clear answer for the correct sample size but a sample size of 30 is held by many researchers to be the minimum number of cases if the researcher plans to use some form of statistical analysis on his/her data. Tuckman (1978) points out that

many researchers select a sample based on the size utilised in a similar study. A sample of 2 235 candidates was used in this study because it was much higher than 383 and the percentage had been used in similar studies as discussed in Chapter II. The sample which represented 2% was preferred in this study because of the findings of Forrest and Vickerman (1982) which were discussed in Chapter II. They found out that if a large group of candidates, representative of the population, wrote examinations in related subjects their mean grades will be the same. Stark (1999) also used a 2% sample when investigating Science standards in Scottish Schools. Random numbers were used for this exercise. Table 8 below shows the population and the sample size for this study. The first column shows the six strata in which the population of students are found in Zimbabwe. The number of candidates in each category is in column two while the last column shows the sample size in each category.

Table 8. Population and Sample Size

School Type	Classification	Population	2% of Population
Government School Former Group A	A	7 227	145
Government School Former Group B	B	31 605	610
Church	C	18 942	385
Independent High Fee Paying Schools	D	5 735	115
Farm & Mine Schools	M	2 800	56
Government & Council Rural Schools	R	45 678	924
	Total	**111 987**	**2 235**

Table 9. below shows the examination centres that were used in the sample. The table also shows the type of centre and the number of candidates that wrote the examinations of Geography 2248 and Science 5006. The first step was to identify the number of candidates needed in the study. The preferred sample was two percent of the population of each category as shown in Table 8. The centres that offered both Geography 2248 and Integrated Science 5006 were listed according to their centre numbers, type and the number of candidates. The type of the centre automatically slotted the centres into their geographical locations. The type of centre also reflects the material and human resources available to the centre as has been described earlier. It was from each of the strata that a random sample of 2% of the candidates was taken. Once the 2% sample within each stratum was achieved, the researcher moved on to the next stratum. It can be noticed from Table 9 that the smallest number of candidates came from Independent high fee paying schools and farm schools. Random numbers were used to pick the sample from the population. The centre numbers that appear on Table 9 are not the actual ones because after the selection was completed the numbers were changed in order to make the identity of the centres anonymous to the general public.

Table 9. Source of the Sample

CENTRE NUMBER	TYPE	NUMBER OF CANDIDATES
11001	A	52
40002	A	93
10003	B	111
10004	B	320
10005	B	230
32006	B	25
86007	B	32
21008	C	151
41009	C	113
89010	C	121
39011	D	115
82012	M	56
20013	R	114
21014	R	143
37015	R	56
40016	R	121
44017	R	92
50018	R	82
50019	R	28
51020	R	108
61021	R	46
73022	R	113
39023	R	80
74024	R	61
91025	R	50
	Total	**2 235**

Research Design

This study used the survey and correlational research designs. Many authors (Cohen & Manion, 1992; Borg & Gall, 1993; Best & Kahn, 1993; etc) have made the point that the correlational research design does not determine whether or not one variable causes the other to change. This study was to establish whether or not there was a relationship between the grades awarded in the same subject and between two subjects over a period of time. The study was intended to determine the degree of consistency in the relationships of grades and the strength of the relationship. Table 10 describes the Best and Kahn (1993) correlation coefficients.

Table 10. Best and Kahn's Interpretation of Pearson's Correlation Coefficient

Correlation Coefficient	Interpretation
.0 to .2	Negligible
.2 to .4	Low
.4 to .6	Moderate
.6 to .8	Substantial
.8 to 1.0	High to very high

The first level of relationship in the table indicates that the relationship that would exist between two variables is of no significance at all while the .2 to .4 would be small. The .4 to .6 correlation coefficient is regarded by Best and Khan (1993) as indicating the existence of a positive relationship between quantities that are compared. In this regard, correlation coefficients in this study were used in order to determine the relationship of grades awarded to candidates. Such relationship studies are referred to by Gall, Borg and Gall (1996) as "aimed primarily at gaining a better

understanding of the complex skills or behavior patterns being studied, and therefore, low correlation coefficients are as meaningful as high coefficients." (p.457).

This point is important because the purpose of this study was to examine the relationship between the grades that were obtained by candidates of the 1996, 1997 and 1998 "O" level Geography and Integrated Science examinations. The investigation of the relationship was for grades obtained by candidates in each subject and the same candidates in the two subjects. Gall, Borg and Gall (1996) further point out that there are many factors that influence behaviour patterns and, therefore, correlation coefficients in the area of educational research are not very high. They state, "correlation in the range of .20 to .40 might be all that one should expect to find for many relationships between variables studied by educational researchers." (p. 459).

The interpretation of the correlation coefficients in this study was in light of the view presented by Gall, Borg and Gall, (1996). The interpretation of the correlation in the view of Best and Kahn (1993) would be from the point of weighing whether a candidate who passes Geography is likely to pass Integrated Science.

Consistency would mean that the standards in the subjects over a period of time were comparable. Lack of consistency would mean that the standards were not comparable.

The Pearson product–moment correlation was the statistical tool that was used to show the measure of correlation in this study. The Pearson's product-moment is used when

the relationship between variables is linear and when the variables are continuous. This is true of the variables that were correlated in this study and hence the use of Pearson's product-moment correlation. It was the data type that led to the use of this method but it is important to note that Cohen, Swerdlik and Phillips (1996) also point out that the Pearson's product–moment correlation is the "most widely used of the several measures of correlation." (p. 131).

The reason for this wide usage is the data type that is linear. In addition, this correlation was used by Forest and Vickerman (1982) when they carried out a subject pairs analysis of subjects that were offered by the JMB examining board.

It is important to point out that the quantitative and qualitative methods are viewed as complimentary in this research. Cohen and Manion (1992) confirm that correlational methodological strategy is quantitative while the survey method is qualitative. The survey method was used to obtain the qualitative aspects of the study. It was important to use the survey method because there was a need to identify standards that were in the question papers and compare them against those that existed in the syllabuses. Information was collected by the use of forms which were filled in by identified experts in the subject areas.

Research Instruments

Three research instruments were used to collect both quantitative and qualitative data from Chief Examiners, markers and Ministry of Education officials. Chief Examiners

are the standard bearers in the Zimbabwe School Examinations Council system. They lead in the question paper development process and are at the forefront of the marking of candidates' scripts. They lead the coordination of marking exercise where markers agree on the marking scheme that is used in the marking of candidates' scripts. They pay particular attention at the point that marks are not transferred from one part question to the other at the stage of marking. If marks are transferred at the marking of scripts stage the marks awarded to candidates would fail to satisfy the requirement of the weightings of the assessment objectives laid out in the syllabus. The Chief Examiners and markers were sent all the question papers which the candidates of the 1996, 1997 and 1998 examinations wrote. They were also supplied with the syllabuses so that they could use them as a reference point when making decisions on the content area and the assessment skills the questions tested. They were asked to indicate on the provided forms the skills which each question tested. The forms also indicated sub-questions so that the judges who were Chief Examiners and markers could pass a decision on each sub question. An example of a question with sub-questions is: 1(a)(i); 1(a)(ii); 1(b); 1(c). This would mean that question 1 had four questions in all. Both Geography and Integrated Science Paper I did not have sub-questions. The judges also indicated on the same forms the content from where each of the questions came. These forms (Appendix J,), the question papers (Appendix L) and syllabuses were sent out and returned by post.

The Examinations Education Officers in the Ministry of Education's nine educational regions were asked to classify schools in their region as per Ministry of Education

classification. These Officers were the appropriate people to do the classification because they were the regional experts in the administration of examinations at a regional level. They knew all the schools in their regions and were the first point of contact in examination matters in the regions. The Ministry of Education appointed men and women with a lot of experience in education to such posts because of the importance of examinations to the Ministry. What the Education Officers provided was cross-checked with the classification that was used at the Zimbabwe School Examinations Council.

Ten judges were also asked to make a categorisation of the content and skills that were in the question papers. They were asked to indicate the content area from the syllabus which they felt each of the questions was coming from (Appendix J). On the classification of questions by skills form, the judges were expected to fill in the skill which each question tested. As pointed out in earlier chapters, each judge used in this study was an expert in one subject area. This means that they were trained teachers and ZIMSEC also trained them in the setting of "O" level question papers, marking of examination scripts, grading and the review of candidates' work. The passing of judgement on whether or not a standard that is indicated in syllabuses is also in question papers, is a task that could not be done by untrained persons. The opinions of lecturers, parents, students or any teacher on the quality of questions were not asked for because these would have been mere opinions on standards and not professional judgements from trained people. The knowledge of subject content in order to teach it does not necessarily mean that that person knows how to identify skills in question papers. If

teachers were already experts in setting examinations and marking them there would have been no reason for examination boards, including ZIMSEC, to train them whenever they were contracted to set and mark examinations.

Summary

A two percent sample was randomly selected from a population of about 110 000 candidates which wrote both the Geography 2248 and Integrated Science 5006 examinations in each of the years 1996, 1997 and 1998. This meant that 2 235 candidates were in the sample. These came from six categories of schools in Zimbabwe, government Former group A, Former group B, Church, High fee paying private, and Farm and Mine. The survey and the correlational research designs were used. The correlational research design was used to determine the existence or non-existence and the strength or weakness of the relationship between standards within each of the subjects and also between them.

CHAPTER IV: DATA COLLECTION AND ANALYSIS

Introduction

This chapter focuses on the sources of data, how the data was collected and analysed. The sources of data discussed in this chapter are the 1996, 1997 and 1998 question papers, the graded candidate lists of the three examinations, and the judgments that were made by experts in the two subject areas. The data is presented in the form of tables and graphs.

Sources of data

The Graded Candidate lists are computer generated. A script mark from each candidate is entered on a marksheet by a marker who would have marked that script. The marker checks the accuracy of the mark entered. The checking system must be put into the context of the marking operation. For every five to seven markers there is one team leader. The responsibility of each team leader is to monitor marking standards by markers in a team. Some of the duties of the team leader in monitoring standards are to see to it that scripts are marked as per instructions in the marking schemes; every page of each script is marked; marks awarded for questions attempted are correctly added; and that total marks on each script are correctly transcribed on to the mark sheet etc. Team leaders in each of the question papers offered by the Zimbabwe School Examinations Council report to the Assistant National Chief Examiner and the National Chief Examiner. These two carry out the same duties as described for the team leaders. The only difference is that they monitor the performance of team leaders. The work of

the Assistant National Chief Examiner is monitored by the Chief Examiner. The Chief Examiner reports to the Subject Officer at the Zimbabwe School Examinations Council.

In a bid to eliminate errors that arise from transcription of marks the Zimbabwe School Examinations Council employed transcription checkers in order to provide a second check on the addition and transcription of marks from scripts to mark sheets. During the time this study was carried out the Council employed checkers after all the marks had arrived at the Council offices.

Optical mark readers at the Zimbabwe School Examinations Council in Harare scan the marks. A panel of experts at the Zimbabwe School Examinations Council in Harare determines grade thresholds for each component and the subject. The Chief Examiner of each question paper plays an important role in the professional decisions on grade cut-off points. By the time the meeting on awards of grades is held the Chief Examiner would have already provided the Council with his/her qualitative comments on the question paper. His/her comments come after consultations with the members of his/her marking team. The process of arriving at recommendations for grade cut-off points by the Chief Examiners has been discussed in Chapter II. After the grading exercise, the Council computer lists every candidate in each subject and grades obtained in each component and subject. This document is known as the Graded Candidate list (Appendix H). The graded candidate lists show the year of the examination, the centre numbers, and the candidate numbers of all candidates who wrote the subject, the numerical grades obtained in each of the components that constitute the syllabus and the

syllabus grade. Where a candidate registered for the examination but failed to write it, an X appears against the candidate number.

The grades on the graded candidate list were entered into Microsoft Excel computer software programme. The Appendix I shows the layout of the data in Microsoft Excel. The columns show the centre and candidate numbers, the subject and paper grades for the years under study. Only a few pages of the output were put in the appendix section because the thesis would have been bulky if all the pages were included.

The Microsoft Excel output also shows the classification of the centres as discussed in Chapter III. This was to ensure that the candidates who were in the sample were from the classification of centres as indicated in Chapter III.

The grades of the same candidate in one year for both Geography 2248 and Integrated Science 5006 were captured into Microsoft Excel computer software programme. This was done for each of the years 1996, 1997 and 1998. Where a candidate wrote only one subject his/her results were not included in the analysis. The study, therefore, analysed the grades of the same candidates who wrote Geography and Integrated Science in a year. It was important to record the performance of the same candidates in the two subjects in a year because, as discussed earlier, the study was to establish whether or not there was a relationship between the grades in the two subjects and also within each of the subjects over the specified period.

The data from the subject experts was collected using a form, Appendix J. Each of the experts was a subject specialist who had taught or was currently teaching the subject at secondary school level. He/she was also a trained and qualified examiner of the subject. As an examiner, he/she set questions, got involved in discussing the questions before the Zimbabwe School Examinations Council accepted them as valid and reliable questions. The discussions are held during the item writing workshops that are organised by the Zimbabwe School Examinations Council. These specialists were also markers of the "O" level scripts. The experts who were used were, therefore, people who had high academic and professional qualifications as well as experience in examinations.

Ten specialists in each of the subjects were sent the 1996, 1997 and 1998 question papers (Appendix L) and syllabuses and asked to classify the questions into the content areas and the skills that they tested candidates who sat for the examinations. Elliot and Massey (1994) and Jones and Lotwick (1979) referred to in Chapter II used six and nine judges respectively in their studies. Ten judges were considered in this study to provide a wide professional opinion on the question papers. However, seven experts in each of the subjects returned their judgements. The three experts who were not involved communicated with me by telephone indicating that they had received the materials but were not in a position to be involved because of heavy commitments. The number of the experts that was used was still considered adequate despite the reduction because the number was still in the range of the numbers that had been used by other researchers referred to earlier. Table 3 (see Chapter II) that assisted the experts in making their

decisions had learning outcomes and verbs that described the learning outcomes. They were also supplied with the assessment objectives from the syllabuses. These assisted them with explanations as to what each assessment objective meant so that they could classify the questions in each question paper appropriately.

The form that the experts filled in for other question papers other than paper one was provided with sub-questions so as to capture the skill that each sub question was testing (see Appendix J). This was important because structured questions are known to have many sub-questions testing different skills.

Some experts indicated that two skills (a lower order and higher order skill) were being tested by the same question. In this regard the higher order skill was the one taken as the skill that the question was meant to test. As shown in Appendix J the experts indicated the skill levels by numbers one to three. The skills are hierarchical in that skill one is easy to achieve while skill three is hard. It is, therefore, not possible to exhibit attainment at skill three without competency at skills one and two. The meaning and weighting of each of these skills was discussed in Chapter II. Appendix J shows the output of the judgements of the experts. Column 1 shows the question number. The other columns show the decisions of each expert on each question and sub-questions. The columns 9 to 11 show the number of judgements under each skill. After column 11 the question number column starts again. It can be noticed that on Geography Paper One, question 1, five judges decided that the question tested the application of skills while two judges decided that the question tested knowledge. In this case the researcher

took the decision of the five judges as that which the question was testing. The decision of the majority of the judges was the one that was used in this study. The judgements of the experts were analysed in Microsoft Excel.

Number of Questions and Skills in Each Paper

It was important to start by analysing the number of questions in each question paper that was answered by the candidates in 1996, 1997 and in 1998. Only one set of question papers was put in as Appendix L to show the reader the structure of the question papers because the inclusion of the question papers of the other two years would have made the appendix section of the thesis bulky. The Zimbabwe School Examinations Council gave the researcher the permission to publish the question papers (Appendix K). The structure of question papers for the other two years remained the same. The analysis of the question paper was done to establish the standard of the question papers over a three-year period. As stated earlier in Chapter One, question papers have a pivotal role in setting standards in an examination system. The standards are reflected in the quality of questions set and the number of questions that are set in an examination from year to year. It would be ridiculous to set nine questions in an examination in one year and the following year set twenty questions because candidates who write those examinations cannot be graded on the same scale. This would mean that the tasks which should be in a question paper would shift from year to year.

The Geography 2248 and Integrated Science 5006 Question Paper Ones were made up of 40 questions each. This number of questions was consistent over the three-year

period. The questions were all multiple choice type and were answered on scanner sheets that were marked by computer. Geography Paper One question papers had sections. These are Mapwork; Physical Geography; Economic Geography; and Population, Settlement, & Trade (see Appendix L). The number of sections and titles remained the same over the three-year period. The Science question papers did not have sectional topics. The structure also remained the same over the three-year period.

The syllabus content and the question papers were compared to investigate the content validity of the question papers over the three-year period. The experts mentioned earlier made the content categorisation. There was agreement among the experts in the content areas that were tested by the 1996, 1997 and 1998 questions papers. The expects were required to complete the content validity form (Appendix J.). They were expected to write the syllabus section as indicated in the syllabus. Tables 11, 13, 15, 17 and 19 are a result of the information from the experts. Columns in these five tables show the following:

Column 1 The year of the examination, 1996, 1997 and 1998.

Column 2 The question numbers as they appeared on the question papers.

Column 3 The syllabus section under which one finds the different topics which candidates were supposed to cover before the examination.

Tables 12, 14, 16, 18, and 20 summarize the information in tables 11, 13, 15, 17, and 19 respectively. The total number of questions that was set under each syllabus topic is

indicated in these summary tables. The comparison of this information and what is stipulated in the syllabuses is referred to in Chapter V. The comparison forms a basis of judging whether or not the standards that were stated in the syllabuses were maintained in the 1996, 1997 and 1998 question papers. In other words, when a syllabus states that a certain number of questions would be set on a certain content area, the question papers should reflect just that so that the standard is maintained.

Table 11. below shows the number of Geography Paper 1 questions that were set on each syllabus topic in each of the years 1996, 1997 and 1998.

Table 11. Syllabus Topics Covered in Geography Paper 1

GEOGRAPHY PAPER 1		
YEAR	**QUESTION**	**SYLLABUS SECTION**
1996	1	Mapwork
1997		Mapwork
1998		Mapwork
1996	2	Mapwork
1997		Mapwork
1998		Mapwork
1996	3	Mapwork
1997		Mapwork
1998		Mapwork
1996	4	Mapwork
1997		Mapwork
1998		Mapwork
1996	5	Mapwork
1997		Mapwork
1998		Mapwork
1996	6	Mapwork
1997		Mapwork
1998		Mapwork
1996	7	Mapwork
1997		Mapwork
1998		Mapwork
1996	8	Mapwork
1997		Mapwork
1998		Mapwork
1996	9	Mapwork
1997		Mapwork
1998		Mapwork
1996	10	Mapwork
1997		Mapwork
1998		Mapwork
1996	11	Mapwork
1997		Mapwork
1998		Mapwork

1996	12	Mapwork
1997		Mapwork
1998		Mapwork
1996	13	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	14	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	15	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	16	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	17	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	18	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	19	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	20	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	21	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	22	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	23	Physical Environment
1997		Physical Environment

1998		Physical Environment
1996	24	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	25	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	26	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	27	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	28	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	29	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	30	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	31	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	32	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	33	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	34	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	35	Population, Settlement, Transport & Trade

1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	36	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	37	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	38	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	39	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	40	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade

Table 12 summaries the number of questions that were set in each of the syllabus sections over the three year period. There is a summary of topics that were in each of the Geography and Integrated Science papers over the years.

Table 12. Summary of the Number of Questions in Geography Paper 1s

Year	Mapwork	Physical Environment	Economic Geography	Population Settlement Transport & Trade
1996	12	13	8	7
1997	12	13	8	7
1998	12	13	8	7

Table 13 shows the number of Integrated Science Paper 1 questions that were set on each syllabus topic in each of the years 1996, 1997 and 1998.

Table 13. Syllabus Topics Covered in Science Paper 1

SCIENCE PAPER 1		
YEAR	**QUESTION**	**SYLLABUS SECTION**
1996	1	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	2	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	3	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	4	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	5	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	6	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	7	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	8	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	9	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	10	Science in Industry
1997		Science in Agriculture
1998		Science in Agriculture
1996	11	Science in Industry

1997		Science in Agriculture
1998		Science in Agriculture
1996	12	Science in Industry
1997		Science in Agriculture
1998		Science in Agriculture
1996	13	Science in Industry
1997		Science in the Community
1998		Science in Agriculture
1996	14	Science in Energy Uses
1997		Science in Industry
1998		Science in Agriculture
1996	15	Science in Industry
1997		Science in Industry
1998		Science in Agriculture
1996	16	Science in Energy Uses
1997		Science in Industry
1998		Science in Agriculture
1996	17	Science in Energy Uses
1997		Science in Industry
1998		Science in Industry
1996	18	Science in Energy Uses
1997		Science in Industry
1998		Science in Industry
1996	19	Science in Energy Uses
1997		Science in Industry
1998		Science in Industry
1996	20	Science in Energy Uses
1997		Science in Industry
1998		Science in Industry
1996	21	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in Industry
1996	22	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in the Community

1996	23	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	24	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	25	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	26	Science in Energy Uses
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	27	Science in Industry
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	28	Science in Structures & Mechanical Systems
1997		Science in Structures & Mechanical Systems
1998		Science in Energy Uses
1996	29	Science in Structures & Mechanical Systems
1997		Science in Structures & Mechanical Systems
1998		Science in Energy Uses
1996	30	Science in Structures & Mechanical Systems
1997		Science in Structures & Mechanical Systems
1998		Science in Structures & Mechanical Systems
1996	31	Science in Structures & Mechanical Systems
1997		Science in Structures & Mechanical Systems
1998		Science in Structures & Mechanical Systems
1996	32	Science in Structures & Mechanical Systems
1997		Science in Structures & Mechanical Systems
1998		Science in Structures & Mechanical Systems
1996	33	Science in the Community
1997		Science in Structures & Mechanical Systems
1998		Science in Structures & Mechanical Systems
1996	34	Science in the Community
1997		Science in the Community
1998		Science in the Community

1996	35	Science in the Community
1997		Science in the Community
1998		Science in the Community
1996	36	Science in the Community
1997		Science in the Community
1998		Science in the Community
1996	37	Science in the Community
1997		Science in the Community
1998		Science in the Community
1996	38	Science in the Community
1997		Science in the Community
1998		Science in the Community
1996	39	Science in the Community
1997		Science in the Community
1998		Science in the Community
1996	40	Science in the Community
1997		Science in the Community
1998		Science in the Community

Table 14. Summary of the Number of Questions in Science Paper 1s

Year	Agriculture	Industry	Energy	Structures	Community
1996	9	6	13	5	8
1997	12	7	7	6	8
1998	16	5	7	4	8

Table 15. Syllabus Topics Covered in Geography Paper 2s

GEOGRAPHY PAPER 2		
YEAR	**QUESTION**	**SYLLABUS SECTION**
1996	1	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	2	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	3	Physical Environment
1997		Physical Environment
1998		Physical Environment
1996	4	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	5	Economic Geography
1997		Economic Geography
1998		Economic Geography
1996	6	Economic Geography
1997		Economic Geography
1998		Economic Geography

1996	7	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	8	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade
1996	9	Population, Settlement, Transport & Trade
1997		Population, Settlement, Transport & Trade
1998		Population, Settlement, Transport & Trade

Table 16. Summary of The Number of Questions in Geography Paper 2s

Year	Physical Environment	Economic Geography	Population Settlement, Transport &Trade
1996	3	3	3
1997	3	3	3
1998	3	3	3

Table 17. Syllabus Topics Covered in Science Paper 2s

SCIENCE PAPER 2		
YEAR	**QUESTION**	**SYLLABUS SECTION**
1996	1	Science in Agriculture
1997		Science in Agriculture
1998		Science in Agriculture
1996	2	Science in Agriculture
1997		Science in Industry
1998		Science in Industry
1996	3	Science in Industry
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	4	Science in Energy Uses
1997		Science in Structures & Mechanical systems
1998		Science in Structures & Mechanical systems
1996	5	Science in Energy Uses
1997		Science in Agriculture
1998		Science in the Community
1996	6	Science in Energy Uses
1997		Science in Agriculture
1998		Science in Agriculture
1996	7	Science in the Community
1997		Science in Industry
1998		Science in Industry
1996	8	Science in the Community
1997		Science in Energy Uses
1998		Science in Energy Uses
1996	9	Science in Agriculture
1997		Science in Structures & Mechanical systems
1998		Science in Structures & Mechanical systems
1996	10	Science in Industry
1997		Science in the Community
1998		Science in the Community
1996	11	Science in Energy Uses

1997		0
1998		0
1996	12	Science in Structures & Mechanical systems
1997		0
1998		0
1996	13	Science in the Community
1997		0
1998		0

Table 18. Summary of The Number of Questions in Science Paper 2s

Year	Agriculture	Industry	Energy	Community	Structures
1996	3	2	4	3	1
1997	3	2	2	1	2
1998	2	2	2	2	2

Table 19. Syllabus Topics Covered in Science Paper 3s

SCIENCE PAPER 3		
YEAR	**QUESTION**	**SYLLABUS SECTION**
1996	1	Science in Industry
1997		Science in Agriculture
1998		Science in Agriculture
1996	2	Science in Agriculture
1997		Science in Industry
1998		Science in Energy Uses
1996	3	Science in Industry
1997		Science in Energy Uses
1998		Science in Industry
1996	4	Science in Energy Uses
1997		Science in the Community
1998		Science in Energy Uses

The name of syllabus section in column three represents a question. Where the content area has not been written down in the table that means that no question from that area was set. Table 20 indicates a zero under that topic where no questions were set.

Table 20. Summary of The Number of Questions in Science Paper 3s

Year	Industry	Agriculture	Energy	Structures	Community
1996	2	1	1	0	0
1997	1	1	1	0	1
1998	1	1	2	0	0

Questions in Science Paper One were set from five syllabus topics over the three-year period. Consistency was observed in the number of questions set from the syllabus topic ***Science in the Community.*** That was a good example of the achievement of comparable standards in the number of questions from a syllabus topic. However, the number of questions from each of the other four topics varied. The biggest variation was observed in 1996 and 1998 when the number of questions in ***Agriculture*** varied by as much as seven. In the same years the number of questions in ***Energy*** varied by six. There were small variations in the number of questions in Science Paper Two. On one hand there was a variation of one question only in ***Agriculture*** and ***Structures*** and on the other there was a variation of two questions in each of the topics, ***Energy*** and ***Community.*** It was interesting enough to note that no question was set on the topic

Structures over the three-year period in Paper Three. Small variations are also observed in this paper.

One way that can be used to determine the quality of questions is by looking at the extent questions confirmed to the specifications of the syllabuses. The importance of assessment objectives in determining the quality of any assessment instrument has been discussed earlier. Appendices to show the requirements of the syllabus in as far as the number of assessment objectives that must be included in a question paper have also been referred to in earlier chapters. It was with this background that experts were used to make some judgements on the quality of the questions that were used in 1996, 1997 and 1998.

The judgements shown in Appendix J were used to generate the data in Table 21. The table shows the analysis of the skills in each question paper. There are six columns in the table. The first column shows the year the question paper was offered to candidates. The second column shows the number of questions where there was no majority decision. The third to the fifth columns show the skills into which the questions were classified. The sixth column shows the total number of questions in each of the examination question paper.

Table 21. Classification Of Questions Into Skills

		Geography Paper 1s			
	No majority decision (0)	Skill level 1	Skill level 2	Skill level 3	Total
1996	1	11	23	5	40
1997	2	16	18	4	40
1998	1	14	17	8	40
		Science Paper 1s			
	No majority decision (0)	Skill level 1	Skill level 2	Skill level 3	Total
1996	0	23	17	0	40
1997	0	24	16	0	40
1998	1	25	14	0	40
		Geography Paper 2s			
	No majority decision (0)	Skill level 1	Skill level 2	Skill level 3	Total
1996	2	14	22	12	50
1997	0	18	22	10	50
1998	1	21	15	10	47
		Science Paper 2s			
	No majority decision (0)	Skill level 1	Skill level 2	Skill level 3	Total
1996	2	26	18	0	46
1997	1	29	11	1	42
1998	0	24	20	0	44
		Science Paper 3s			
	No majority decision (0)	Skill level 1	Skill level 2	Skill level 3	Total
1996	2	8	8	16	34
1997	2	2	5	5	14
1998	2	1	7	9	19

The number of questions where judges failed to have a majority decision is very small. For example, out of 40 questions in the 1996 Geography Paper One there was only one question on which there was no majority decision by the judges. The questions where

there was no majority decision were not used in the analysis. The total number of such questions was so small that it did not distort the overall picture.

The number of questions under each skill in each paper was expressed as a percentage (see Table 22). This provided a platform for comparison because in some papers the total number of questions differed from year to year.

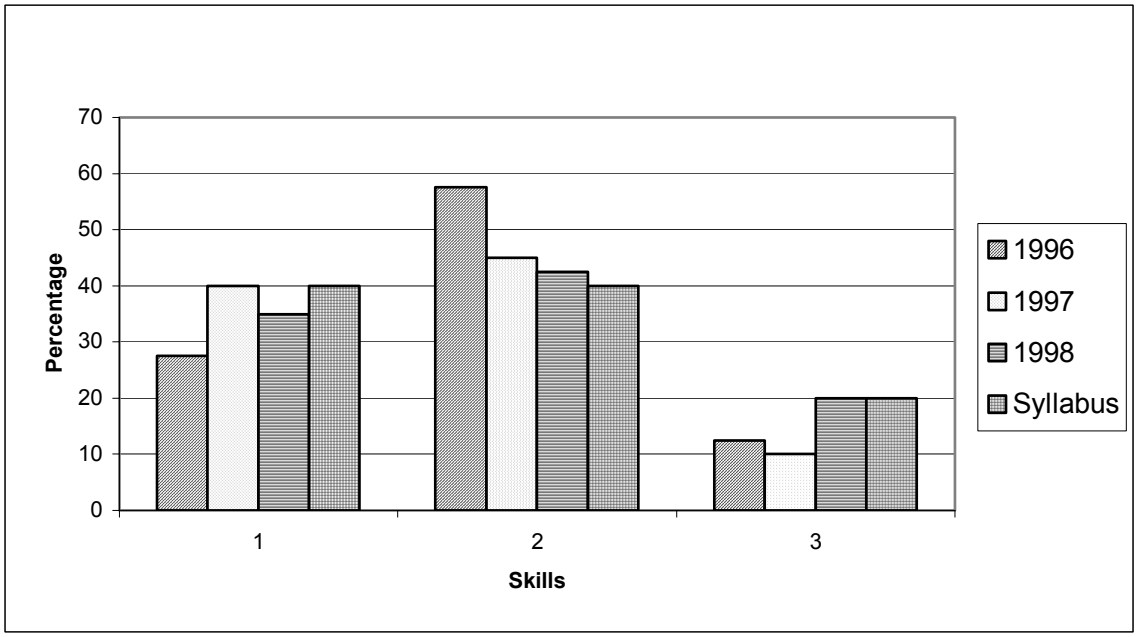
Table 22. The Percentage of Skills in The Geography and Science Papers

	Geography Paper 1s		
	Skill level 1	Skill level 2	Skill level 3
1996	27.5	57.5	12.5
1997	40.0	45.0	10.0
1998	35.0	42.5	20.0
	Science Paper 1s		
	Skill level 1	Skill level 2	Skill level 3
1996	57.5	42.5	0.0
1997	60.0	40.0	0.0
1998	62.5	35.0	0.0
	Geography Paper 2s		
	Skill level 1	Skill level 2	Skill level 3
1996	28.0	44.0	24.0
1997	36.0	44.0	20.0
1998	44.7	31.9	21.3
	Science Paper 2s		
	Skill level 1	Skill level 2	Skill level 3
1996	56.5	39.1	0.0
1997	69.0	26.2	2.4
1998	54.5	45.5	0.0
	Science Paper 3s		
	Skill level 1	Skill level 2	Skill level 3
1996	23.5	23.5	47.1
1997	14.3	35.7	35.7
1998	5.3	36.8	47.4

The data in Table 22 was used to produce the bar graphs shown is figs. 1 to 5.

It can be noticed in Figure 1. that the percentage of skills at each level was not exactly the same over the three years, 1996, 1997 and 1998. The percentage of skills which are indicated in the syllabus for this paper is also shown in Figure.1.

FIGURE 1. SKILLS IN GEOGRAPHY PAPER 1



The question papers that were set over the period under study did not strictly adhere to this requirement but the relationship with the syllabus prescription was fairly impressive. It is important to hasten to add that although the weighting of skills must be adhered to in order to achieve comparability, strict adherence to the stated percentages is not, as stated in the syllabus document, a requirement. The Geography syllabus

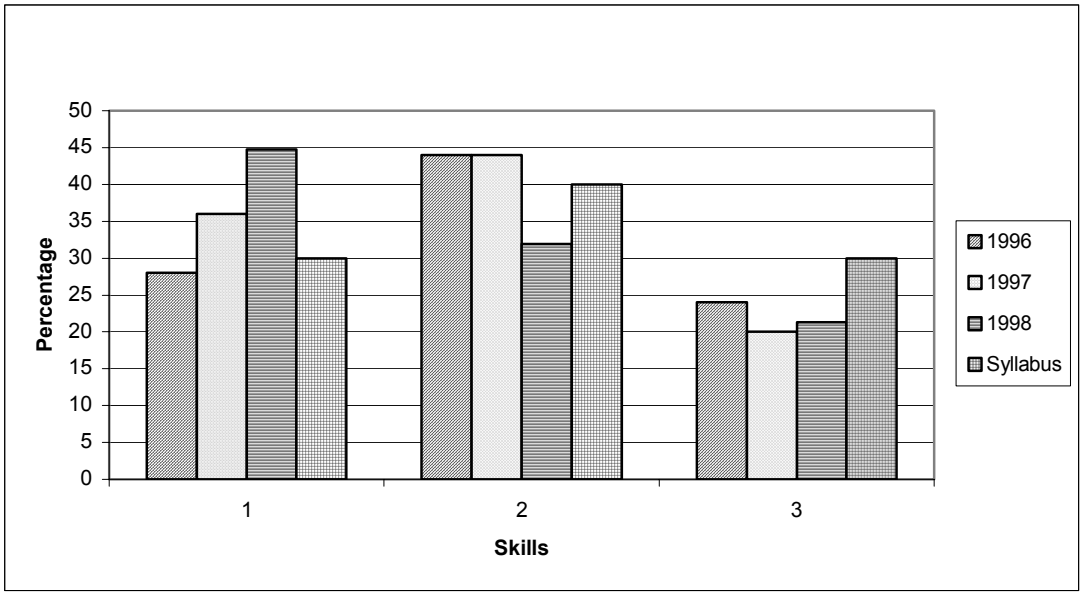
document (1995) states, "The assessment objectives are weighted to give an indication of their relative importance. They are not intended to provide a precise statement of the number of questions or marks allocated to particular objectives." p23.

It can be argued, however, that although the syllabus makes this point, the number of questions with the same number of skills should not be significantly different from one year to the next otherwise the standard set in one year could be significantly different from other years. It is important to bear in mind that although the syllabus document refers to variations, the document does not quantify the variation. This is, therefore, a weakness of the syllabus because without quantifying the variations that should be permitted it would become difficult to know the reasonable variations that can be tolerated in an examination from one year to the next.

The standard that was set in the Geography question papers compare very well at Skill 1 for the 1997 and the 1998 papers; at Skill 2 for the 1997 and 1998 papers and at Skill 3 for the 1996 and 1997 papers. Questions at Skill 1 in 1996 were less than in any other year and so were the questions at Skill 3 for the 1997 question paper. Skill 2 for the 1996 paper was way above the required 40% weighting. It was at 57.5%. In general, most questions in Geography Paper One were pitched at Skill 2. In the three years the questions at this level of difficulty constituted a percentage higher than the 40% requirement. In 1996 and 1997 the percentage of Skill 3 questions was below the requirement.

Geography Paper Two had nine questions in each of the examinations that were offered in the three years. There was consistency here. These questions had sub-questions but the number of sub-questions of each of the nine questions differed from year to year. The 1996 and 1997 examinations had fifty sub-questions while the 1998 examination had forty-seven sub questions. The last column of Table 21 shows the total number of all questions in each paper. The syllabus requires that the percentage of Skill 1 and 3 be 30% each while Skill 2 is 40%. The 1998 question paper stood out of the pattern which the other two examinations showed in as far as the skill distribution was concerned (Figure 2.).

FIGURE 2. SKILLS IN GEOGRAPHY PAPER 2



The number of questions for Skill 1 in 1998 was more than those for Skill 2. This should have made the paper slightly easier than the question papers for 1996 and

1997. It is also important to note that it was at Skill 2 in 1998 where there was a reduction of questions when compared to the other years. At Skill 3 the questions were even more than those in the 1997 question paper. The graph (Fig. 2.) shows the general pattern that for the three years, setters provided less than the stipulated questions with Skill 3 while questions at Skill 2 were more than the required in 1996 and 1998. At Skill 1 the questions were more in 1997 and 1998 than the 30% requirement. These variations had a serious implication on comparability of standards in the two subjects. The variations can be seen on the graph as one compares each of the years 1996, 1997 and 1998 with the syllabus requirement. The number of tasks, as represented by the number of sub-questions in each question paper, gave candidates in the different years different amount of work at examination time.

The Integrated Science Paper One had forty multiple-choice questions. The number of questions did not vary from year to year. The syllabus document pointed out that each paper covered the whole syllabus but the paper did not have sections. Use of syllabus sections in question papers is a method used in the process of setting question papers to ensure that there is adequate content coverage in examinations. This, of course, would have a bearing on comparability of standards in that the question papers would carry the same domains from one year to the next. However, Paper Two had two sections. Section A had short-answer and structured questions and Section B had five questions that were to be answered in prose form. Candidates were also expected to label diagrams in this examination paper. The number of sub-questions in Section A varied from one year to the next. In 1996 candidates were asked to answer a total of 26

questions, in 1997 there were 22 questions and in 1998 there were 25 questions. However, in Section B the number of tasks was the same over the three-year period.
The Integrated Science syllabus document (1995) stated that Paper Three, the written alternative to practical test is " a written paper of compulsory short-answer or structured questions ….. designed to test familiarity with laboratory practical procedures." p6.

The total number of questions that were asked in each of the years under study remained at four. However, the total number of questions including sub-questions varied greatly. In 1996 there were 34, in 1997 there were 14 and in 1998 there were 19. This variance could have negative impact on examination standards because it was large. The 1996 candidates were subjected to more than double the number of tasks than the 1997 candidates. Not only did the number of questions differ in these years but so did the skills that were tested. Table 22 shows the differences in the number of the skills that were tested in the question papers.

The distribution of skills in Science Paper One showed a different pattern from that of Geography Paper One. This was a reflection of the requirement of the two different syllabuses. The experts did not find any question that was testing Skill 3 in Science paper One. This study has revealed that the percentage of Skill 1 in this paper varied from 57.5 to 62.5 yet the syllabus stated 70. Skill 2 varied from 35% to 45.5% yet the syllabus stated 30%. Skill 3, though it is not weighted in the syllabus, was found in paper 2 and 3. The number of questions that tested Skill 1 over the three-year period is comparable. A difference of 5% was registered at Skill 2 between 1997 and 1998.

FIGURE 3. SKILLS IN SCIENCE PAPER 1

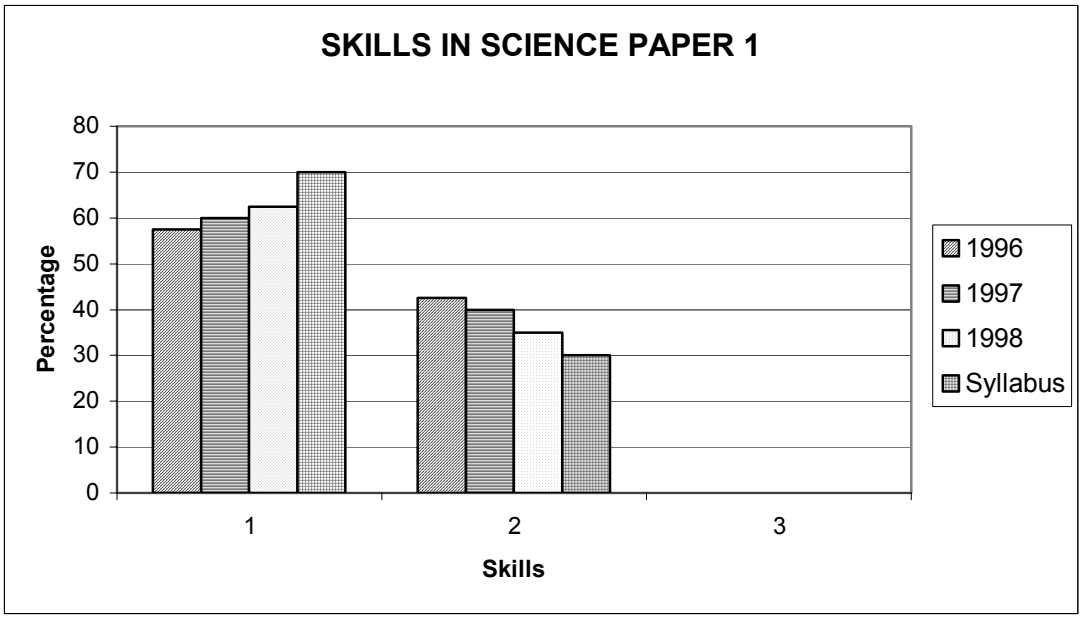
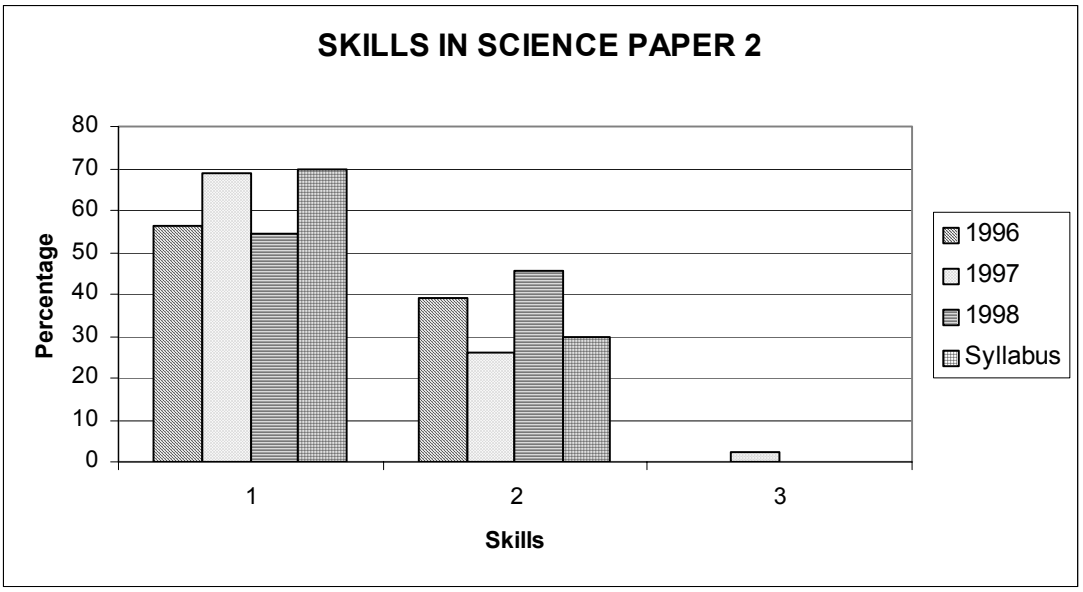


FIGURE 4. SKILLS IN SCIENCE PAPER 2

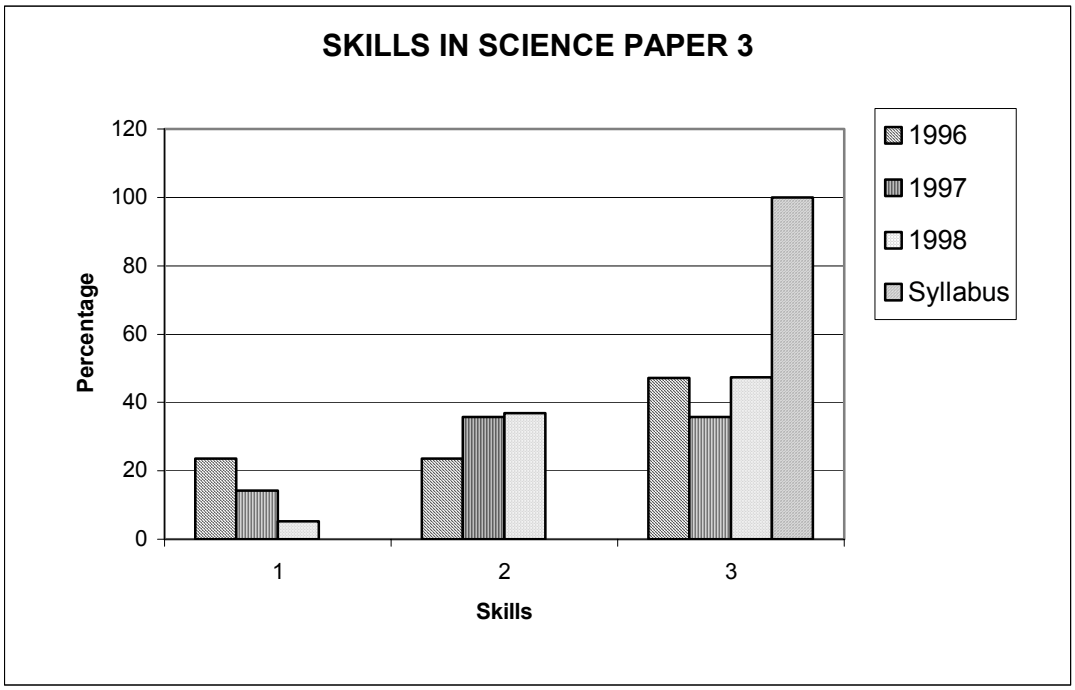


The total number of questions (including sub-questions) which candidates were expected to answer in Science Paper Two differed from year to year. In 1996 there were 46; in 1997 there were 42; and in 1998 there were 44. Figure 4 shows a very interesting pattern. It was only in 1997 that there was a question that tested Skill 3. The paper had a very high number of Skill 1 questions in 1997. The distribution of skills in question papers was comparable in 1996 and 1998. It can be concluded that the 1996 and the 1998 Science Question papers failed to stretch candidates as the 1997 question paper did.

The number of Science Paper Three questions, like those of Science Paper Two, varied from year to year. It can be noticed in Table 21 and Figure 5 that in 1996 there were 34

questions; in 1997 there were 14; and in 1998 there were 19. The distribution of skills in Science Paper Three clearly showed that it is a practical paper in that the number of questions on skill levels 2 and 3 were more than those for skill level 1. Except for 1997 when the percentage of questions at skill 2 and skill 3 was the same, the other two years had a higher number of questions pitched at Skill 3.

FIGURE 5. SKILLS IN SCIENCE PAPER 3



The skills that were tested in the subjects of Geography and Integrated Science were compared. It is logical to say that if there are more higher order skills in a question paper that question paper and subject is deemed to be more difficult than one that has less. Tables 21 and 22 show the number of skills and the percentage of these to the total. Both subjects clearly showed variations but Integrated Science Paper 3 seemed

to have a bigger difference between syllabus requirements and what was actually in the question papers.

Performance of Candidates in The Two Subjects

It was pointed out in Chapter II that the performance of candidates in a subject is reflective of a standard so set. The other conditions that affect the performance of candidates were referred to in Chapter II. These are the calibre of students from one year to the next at the same school and the quality of instruction. Arguments were presented on these two points in Chapter II. The grades that were achieved by candidates in each of the papers of the subjects under study were put in Microsoft Excel computer software programme as pointed out at the outset of this chapter. Subject mean grades and paper mean grades were calculated and Table 23 shows the mean grades in the two subjects under study.

Table 23. Mean and Mode Grades for Geography and Science

Geog												
	1996 Syll	1996 P1	1996 P2		1997 Syll	1997 P1	1997 P2		1998 Syll	1998 P1	1998 P2	
Mean Grade	6.1	6.1	5.9		6.5	5.9	6.7		6.0	5.7	6.0	
Mode	9	9	9		9	7	9		9	7	9	
Std. Dev	2.5	2.4	2.6		2.3	2.4	2.4		2.4	2.3	2.5	
Science												
	1996 Syll	1996 P1	1996 P2	1996 P3	1997 Syll	1997 P1	1997 P2	1997 P3	1998 Syll	1998 P1	1998 P2	1998 P3
Mean Grade	5.7	5.2	5.9	5.7	5.7	4.4	5.9	5.6	5.5	4.4	5.7	5.8
Mode	7	1	9	9	7	1	9	7	8	1	9	9
Std. Dev	2.8	2.9	2.7	2.8	2.7	2.9	2.6	2.6	2.7	2.9	2.7	2.6

Syll - Syllabus
P1 - Paper 1
P2 - Paper 2
P3 - Paper 3

In order for the reader to interpret the numerical grades that are used in the analyses, it is important to indicate how they are represented in alpha on the Ordinary Level certificate. The numerical grades are represented by alpha grades as follows:

Numerical Grade	**Alpha Grade**
1	Upper A
2	Lower A
3	Upper B
4	Lower B
5	Upper C
6	Lower C
7	D
8	E
9	U

One must understand that the bigger the value of a numerical grade, the poorer the grade.

The 1998 Geography examination proved easiest of the three Geography examinations. Whereas in 1998 the mean grade was 6.0 in 1996 and 1997 the mean grades were 6.1 and 6.5 respectively. In terms of alpha grades the mean grade in Geography over the three-year period was a C. It can also be noticed that the mode grade over the three-year period was 9, that is grade U, signifying poor performance because many

candidates failed to achieve passing grades. The relevant aspect of this analysis is the nature of the relationship of the grades that were awarded to candidates from one year to the next. In the case of the subject of Geography the analysis illustrates that the mean grade was the same and so was the mode. The syllabus grades were marginally more spread in 1996 than in 1998 and 1997 at standard deviation of 2.5, 2.4 and 2.3 respectively. These facts illustrate a positive relationship within this subject over a three-year period. The aggregation effect distorts the differences that were noted in the paper grades. Candidates achieved higher grades in the paper 1s of 1997 and 1998. In 1996 paper 2 was easier than paper 1. What is communicated to candidates is the overall achievement in a subject and so the differences that emerge in the paper grades though noted yielded a mean grade that is the same for the syllabus over a three-year period.

The mean grades in Science also show the same consistency in the performance of the candidates as shown in the subject of Geography. The 1997 and the 1998 paper ones proved easier to candidates than other question papers. It can be noticed that the same mean grade of 4.4 was achieved in 1997 and 1998 in Paper 1. This is grade B in alpha. From the point of view of the grades being achieved by candidates, this represents similar standards in the two examinations. The 1996 and 1997 paper two proved to be more difficult than the 1998 one. The difference is really marginally because they would all fall in the same alpha grade of C. Paper ones were the easiest of the three papers. This can be noticed by the mode of the grades. While the mode of Paper one was 1, it was 7 and 9 in the other two papers. The reader needs to be reminded of the

fact that this indicates similarity of standards in each of the papers within this subject over three years. The performance of candidates agrees with the judgements that were made by experts on the skills that the question papers for these years were testing. It can be noticed on Fig 3 (Skills in Science Paper 1) that in 1997 and 1998 there were more questions with the lower order Skill 1 than in 1996. The most common grade in all the Science Paper Ones (Grade 1) supports the fact that the question papers tested low order skills.

The 1996 and 1997 Integrated Science Paper 2 were more difficult than the 1998 paper. The mean grade was 5.9. This is however marginally higher than the 1998 mean at 5.7. Science Paper 3 of 1998 was another difficult paper with a mean grade of 5.8 because the question paper had the lowest number of questions at Skill 1 than the question papers of the other two years. Furthermore, there were many questions at Skill 3 in the question paper. This was revealed by the experts who categorised the skills of these papers (Table 21).

The mean grades showed clearly that the Geography mean grade of C that was achieved over the three-year period was an inferior C when it is compared to the mean grade of C in Integrated Science. Table 23 has illustrated that the standards within each of the subjects were the same because the mean grades were the same and also that though the mean grades of the two subjects were numerically different, they all fall into the Grade C category. This shows that the standards were

comparable within each of the subjects and between the subjects in the period 1996 to 1998.

The relationship between the grades was also examined in terms of the correlation coefficients. The Pearson's product-moment correlation that was discussed in Chapter III was used. Tables 24 and 25 show these correlation coefficients. The correlation coefficient between the performance of candidates in Geography papers was the same for the three years at 0.4.

Table 24. Relationship of Candidates' Performance in Geography

Error! Not a valid link.

Geography Syllabus, Paper and Year of examination	**Correlation coefficient**
Syllabus 1996 and 1997	+ 0.4
Syllabus 1997 and 1998	+ 0.5
Syllabus 1996 and 1998	+ 0.5
Paper 1 1996 and 1997	+ 0.4
Paper 1 1997 and 1998	+ 0.4
Paper 1 1996 and 1998	+ 0.4
Paper 2 1996 and 1997	+ 0.4
Paper 2 1997 and 1998	+ 0.4
Paper 2 1996 and 1998	+ 0.4

Table 25. Relationship of Candidates' Performance in Integrated Science

Science Syllabus, Paper and Year of examination	Correlation coefficient
Syllabus 1996 and 1997	+ 0.5
Syllabus 1997 and 1998	+ 0.5
Syllabus 1996 and 1998	+ 0.4
Paper 1 1996 and 1997	+ 0.3
Paper 1 1997 and 1998	+ 0.3
Paper 1 1996 and 1998	+ 0.3
Paper 2 1996 and 1997	+ 0.4
Paper 2 1997 and 1998	+ 0.5
Paper 2 1996 and 1998	+ 0.4
Paper 3 1996 and 1997	+ 0.4
Paper 3 1997 and 1998	+ 0.5
Paper 3 1996 and 1998	+ 0.4

The correlation coefficients between the 1997 and 1998 and also between the 1996 and the 1998 Geography syllabus grades were 0.5. This was higher than the syllabus grades of 1996 and 1997 at 0.4. The correlation coefficient of the performance of each of the papers was consistent at 0.4. The correlation coefficients for paper ones was the same as for paper twos. According to the table given in Chapter III on the interpretation of correlation coefficients the standards that were in the papers and syllabuses was moderate. However, the correlation coefficients must be interpreted in the light of the argument of Gall, Borg, and Gall (1996) presented in Chapter III. They argue that from the point of view of educational research, the relationships between the performance of candidates between question papers and subjects is significant when

it lies between 0.2 and 0.4. This was exceeded in the relationship between the syllabus grades of the 1997 and 1998 examinations.

The lowest correlation coefficients were recorded between all the Science Paper Ones at 0.3. The highest correlation coefficients were recorded between the Science syllabuses of 1997 and 1998; 1996 and 1997; and also between Science Paper Twos of 1997 and 1998; Paper Threes of 1997 and 1998.

The correlation coefficients between the performance of candidates in the syllabuses were also compared. It has to be remembered that the performance of candidates that was compared was of the same cohort of students in each year that wrote the examinations in the two syllabuses. As shown in the Table 26, the relationship was substantial in 1996 and moderate in 1997 and 1998. These correlation coefficients were significant at 0.01 level. However, the important point to raise here is that there is a positive relationship between the performances of candidates in the two subjects. Stakeholders expect the existence of a positive relationship between grades awarded to candidates.

The similarity of standards is obviously higher at 0.5 than 0.4 but this study buttresses its judgement on the views of Gall, Borg and Gall (1996) stated in Chapter III. The correlation coefficient in the range of .2 and .4 indicate presence of good relationship in educational research. The correlation coefficients indicate that there was a positive relationship which is close to each other. This is important.

Table 26. Relationship of Candidates' Performance in Geography and Integrated Science

Year	Correlation coefficient of the Geography and Science Syllabus Grades
1996	+ 0.7
1997	+ 0.5
1998	+ 0.6

It is important to note that the relationship that is sought by this study was whether or not the same candidates who sat for the two subjects were getting similar grades. The correlation coefficients show that this was the case indicating that the standards that were set in each of the years were similar.

The number of candidates in the sample that achieved particular grades was compared to establish whether or not this number was comparable from year to year. The number of candidates was expressed as a percentage at each grade. It has been established in earlier chapters that the characteristics of a population of students that attend a particular school does not, in general, change from year to year. The students are primarily drawn from the same neighbourhood every year. That being the case the performance of the students should not be significantly different from one year to the next.

Table 27 below shows the number of Integrated Science candidates in the sample at each of the grades 1 to 9. Table 28 has the same information as Table 27 except that the number of candidates at each grade has been expressed as a percentage of the total number of candidates in the sample. The same has been done for Tables 29 and 30 for Geography.

Table 27. Number Of Candidates At Each Grade In Science Papers

Grades	1996 Syll.	1996 P.1	1996 P.2	1996 P.3	1997 Syll.	1997 P.1	1997 P.2	1997 P.3	1998 Syll.	1998 P.1	1998 P.2	1998 P.3
1	327	509	231	300	267	705	242	254	313	687	274	170
2	100	105	140	139	113	133	73	110	110	180	111	117
3	136	96	145	146	158	131	185	157	209	129	219	304
4	166	184	186	173	154	251	133	187	149	103	146	180
5	164	150	187	159	232	148	249	207	205	222	224	157
6	147	140	172	137	165	21	206	247	172	103	161	188
Total No. at Grades 1 to 6	1040	1184	1061	1054	1089	1389	1088	1162	1158	1424	1135	1116
7	432	417	371	413	438	426	386	521	377	426	378	349
8	389	431	349	347	338	234	318	215	397	298	315	346
9	374	203	454	421	370	186	443	337	303	87	407	424
Total No. at Grades 7 to 9	1195	1051	1174	1181	1146	846	1147	1073	1077	811	1100	1119
Sample total	2235	2235	2235	2235	2235	2235	2235	2235	2235	2235	2235	2235

Table 28. Percentage Of Candidates At Each Grade In Science Papers

Grades	1996 Syll.	1996 P.1	1996 P.2	1996 P.3	1997 Syll.	1997 P.1	1997 P.2	1997 P.3	1998 Syll.	1998 P.1	1998 P.2	1998 P.3
1	14.6	22.8	10.3	13.4	11.9	31.5	10.8	11.4	14.0	30.7	12.3	7.6
2	4.5	4.7	6.3	6.2	5.1	6.0	3.3	4.9	4.9	8.1	5.0	5.2
3	6.1	4.3	6.5	6.5	7.1	5.9	8.3	7.0	9.4	5.8	9.8	13.6
4	7.4	8.2	8.3	7.7	6.9	11.2	6.0	8.4	6.7	4.6	6.5	8.1
5	7.3	6.7	8.4	7.1	10.4	6.6	11.1	9.3	9.2	9.9	10.0	7.0
6	6.6	6.3	7.7	6.1	7.4	0.9	9.2	11.1	7.7	4.6	7.2	8.4
% Passing	46.5	53.0	47.5	47.1	48.8	62.1	48.7	52.1	51.9	63.7	50.8	49.9
7	19.3	18.7	16.6	18.5	19.6	19.1	17.3	23.3	16.9	19.1	16.9	15.6
8	17.4	19.3	15.6	15.5	15.1	10.5	14.2	9.6	17.8	13.3	14.1	15.5
9	16.7	9.1	20.3	18.8	16.6	8.3	19.8	15.1	13.6	3.9	18.2	19.0
% Failing	53.4	47.1	52.5	52.8	51.3	37.9	51.3	48.0	48.3	36.3	49.2	50.1

The 1997 Integrated Science Question Paper 1 was the easiest of the Science papers over the three-year period. About thirty-one percent (31.5%) of the candidates in the sample achieved grade 1. Candidates performed poorly in the 1996 and 1997 Paper Twos at grade 1.

Table 29. Number of Candidates at Each Grade in Geography Papers

Grades	1996 Syll	1996 P. 1	1996 P.2	1997 Syll	1997 P. 1	1997 P.2	1998 Syll	1998 P. 1	1998 P.2
1	94	64	186	84	106	119	106	121	172
2	123	136	116	92	149	62	118	149	119
3	219	298	161	140	235	117	194	197	152
4	174	101	198	161	205	144	210	232	195
5	217	217	230	193	230	185	232	257	215
6	255	227	207	229	213	197	275	270	225
Total No. at Grades 1 to 6	**1082**	**1043**	**1098**	**899**	**1138**	**824**	**1135**	**1226**	**1078**
7	310	436	518	394	438	372	354	427	342
8	317	304	40	367	261	332	328	295	299
9	526	452	579	575	398	707	418	287	516
Total No. at Grades 7 to 9	**1153**	**1192**	**1137**	**1336**	**1097**	**1411**	**1100**	**1009**	**1157**
Total of Sample	2235	2235	2235	2235	2235	2235	2235	2235	2235

Table 30. Percentage Of Candidates At Each Grade In Geography Papers

Grades	1996 Syll	1996 P. 1	1996 P.2	1997 Syll	1997 P. 1	1997 P.2	1998 Syll	1998 P. 1	1998 P.2
1	4.2	2.9	8.3	3.8	4.7	5.3	4.7	5.4	7.7
2	5.5	6.1	5.2	4.1	6.7	2.8	5.3	6.7	5.3
3	9.8	13.3	7.2	6.3	10.5	5.2	8.7	8.8	6.8
4	7.8	4.5	8.9	7.2	9.2	6.4	9.4	10.4	8.7
5	9.7	9.7	10.3	8.6	10.3	8.3	10.4	11.5	9.6
6	11.4	10.2	9.3	10.2	9.5	8.8	12.3	12.1	10.1
% Passing	48.4	46.7	49.2	40.2	50.9	36.8	50.8	54.9	48.2
7	13.9	19.5	23.2	17.6	19.6	16.6	15.8	19.1	15.3
8	14.2	13.6	1.8	16.4	11.7	14.9	14.7	13.2	13.4
9	23.5	20.2	25.9	25.7	17.8	31.6	18.7	12.8	23.1
% Failing	51.6	53.3	50.9	59.7	49.9	63.1	49.2	45.1	51.8

The distribution of the grades in the subject of Geography was also of interest. More candidates passed with grade 3 in 1996 in the Geography syllabus than in the other two years when the percentage pass rate was 8. The 1996 Paper 1 had very few candidates who achieved grade 1 when compared to the other two years. This is supported by the difficulty of the question paper as shown on Table 22 which indicates that the paper had the least number of question with the lowest order Skill 1 and had the highest number of skills at level 2. In other words, the standard of the question paper was reflected in the performance of candidates.

What is of significance is the trend that was set in the three-year period. The percentage of candidates rose steadily from grade 1 to grade 9 except at grade 8 where the number fell in the three years. This was true for the three years. More candidates achieved grade 3 in 1996 than in 1997 and 1998. The trends shown over the three-year period are indicative of a similar standard that was applied in the three examinations.

The performance of candidates in Geography Paper 1 presented a different pattern from that of Paper 2. Except for the year 1996 when the performance was low at grade 1 and 4 for Paper 1, candidates produced better grades in this paper than in Geography Paper 2. In Paper 1, the percentage of candidates that achieved grades 3, 7 and 9 was much more than at other grades in 1996. The same can be said for the number of candidates at grade 7 for 1998. The performance in Paper 2 was generally the same over the years except for 1997 when the percentage of candidates at grades 1 to 6 was lower than in other years. The percentage of candidates who got grade 7 in 1996 was way out of the trend. The spread of candidates at grades was generally the same in the two papers with the highest percentage of candidates being at grades 7 and 9. However, the percentage of candidates at these two grades was lower in Paper 1 than in Paper 2 but the trend is similar. This can generally be interpreted to mean that the two papers were performing in the same way. It needs to be remembered that the evidence that showed that the two question papers systematically sampled the syllabus content has already been presented. This could well be the result of how the standard was carried from one year to the next.

The performance of candidates in Science presented a different picture from that of Geography.

The candidates who failed to achieve grade 1 in Geography managed to achieve that grade in Science. It can be observed that the number of candidates who achieved grades 1 and 9 showed a steady increase as the grades became poorer in Geography than in Science. The number of candidates at grades 3 and 8 did not quite fit into the pattern. Science Paper 1 was the easiest of the Science papers. Between 22.8% and 31.5% of the candidates were achieving grade 1 in Paper 1 while in Paper 2 they were about 10% and in Paper 3 they were between 7.6% and 13.4%. The trends that were set over the three-year period were not very different in the Science papers. The performance that was way out of others was that of 1997 Paper 1 at grade 1.

The grades that were achieved by the candidates that wrote the Geography 2248 and Integrated Science 5006 subjects were put on box and whisker plots to establish their spread. This was done to show graphically the relationship of the grades. Figures 6 to 8 show the plots

FIGURE 6. BOX AND WHISKER PLOT FOR THE SCIENCE GRADES

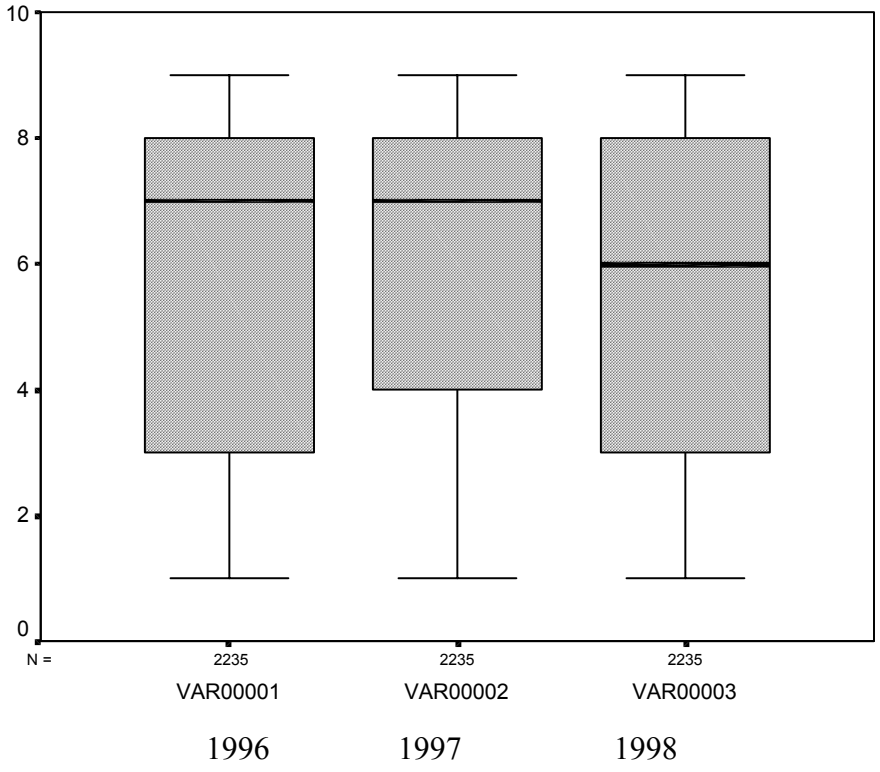
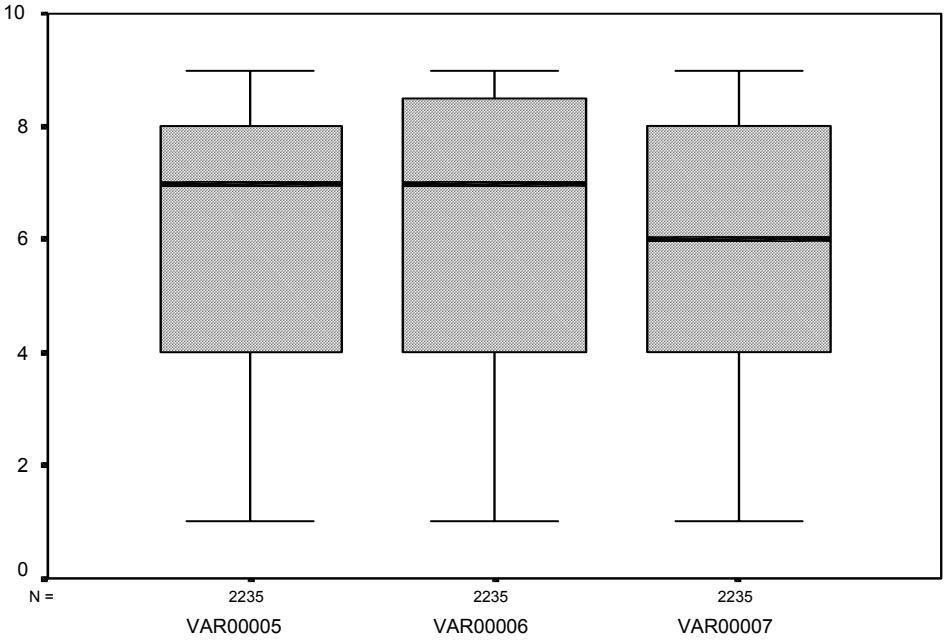
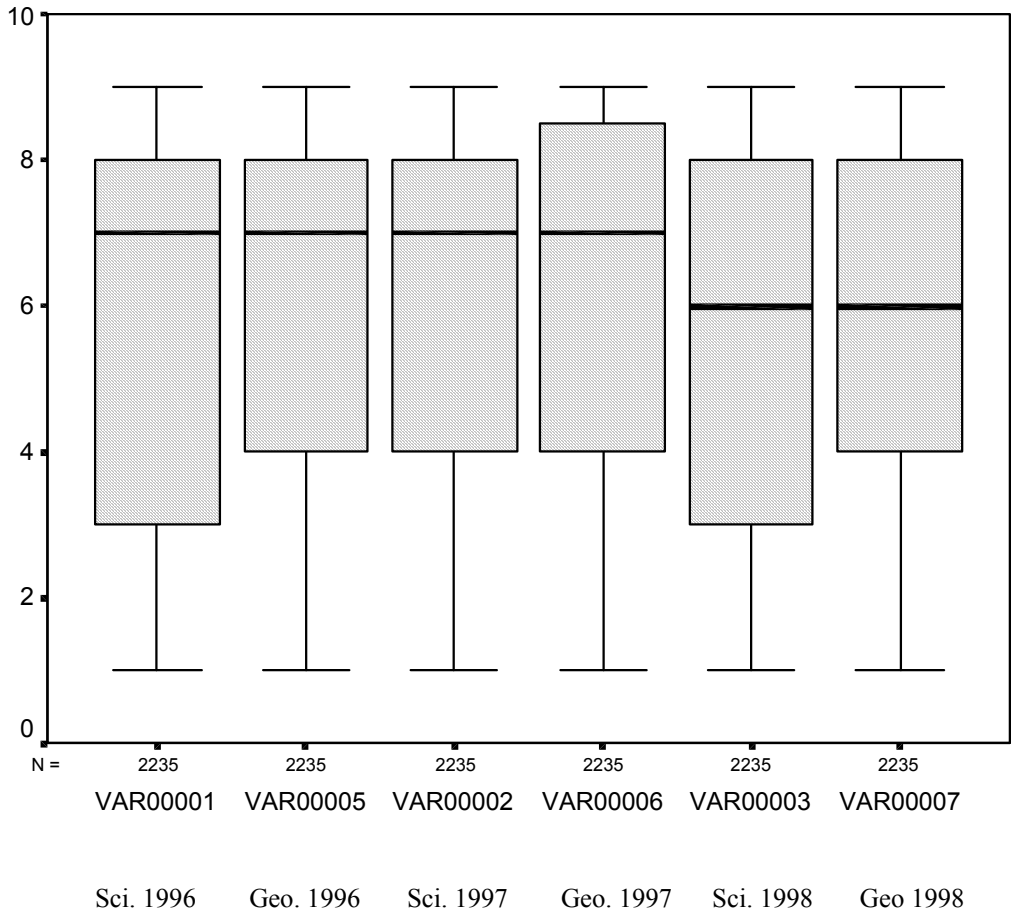


FIGURE 7. BOX AND WHISKER PLOT FOR THE GEOGRAPHY GRADES



1996 1997 1998

FIGURE 8. BOX AND WHISKER PLOT FOR THE SCIENCE AND GEOGRAPHY GRADES



It can be noticed that though the plots for all the subjects were not all exactly the same, considerable similarity can be observed. Except for the 1998 medians for both Geography and Integrated Science the others were similar. The 1998 medians for the two syllabuses were, however, the same. The inter quartile ranges were about the same at 4 and 5. Further to this the spread on the grades around the mean were the same as shown by the standard deviations in the Table 23 above.

Summary

This chapter has dwelt on the quality of the question papers that were given to the 1996, 1997 and 1998 candidates in Geography and Integrated Science. The judgements that were made by experts in testing were discussed and the data collected from these people was presented. The performance of the 2 235 candidates in the sample was also analysed in this chapter.

CHAPTER V: DISCUSSION AND RECOMMENDATIONS

Introduction

This study set out to investigate whether or not examination standards set in Geography 2248 and Integrated Science 5006 at the GCE Ordinary level in Zimbabwe existed. The examinations were first offered in the country in 1990. Before 1990 overseas boards such as the University of Cambridge Local Examinations Syndicate, the Associated Examining Board and the University of London School Examinations Board set Geography and Science examinations for candidates in the country. The investigation sought to establish whether or not there was similarity of standards in two ways. These are within each of the subjects over the three-year period and between the two subjects over the same period of time. The study also set out to provide stakeholders with an approach of interpreting the grades that were awarded to candidates who took the examinations so that the attainment of candidates in different subjects could be understood by the stakeholders in terms of the assessment objectives upon which question papers were based. The review of literature established that a study of this kind had not been carried out in the history of the localised Zimbabwean syllabuses and so this was the first time that information in this area had been gathered and analysed. From this point of view the study gives stakeholders in the area of educational assessment in Zimbabwe new knowledge. Discussion in this chapter is centred on the evidence that the study gathered on the degree of similarity in the examination standards and the research questions that were raised in Chapter I became an important point of reference in the discussion and recommendations made. The three research questions are encapsulated in the sub-headings of this chapter. The evidence presented

shows that question setters in both subjects used the benchmarks for standards, the assessment objectives that are in the syllabuses. However, the standard over the three-year period was better achieved in the subject of Geography 2248 than in Integrated Science 5006 because Geography 2248 had higher content validity than Integrated Science 5006. The evidence gathered shows that the question papers set in Geography 2248 reflected the demands of the syllabus much more than the question papers that were set in Integrated Science 5006. This, therefore, can be interpreted to mean that the Geography 2248 question setters over the three-year period were better trained than those of Integrated Science 5006. Looking at the same factor differently one could state that if the Integrated Science 5006 setters received the same training as the Geography 2248 setters then they were not effective in the discharge of their duty of setting question papers with the required content validity. The instructions for question setters in the syllabus document for Geography are better put across than those of Integrated Science. The Integrated Science 5006 instructions give the setter a free hand that results in a negative impact on the maintenance of standards. Instructions to question setters in syllabuses for the same examination must be the same so that there is comparability of standards across subjects.

The study recommends a method of reporting candidates' performance that is based on the assessment objectives which were found to be a basis of comparing standards set in Geography 2248 and Integrated Science 5006. Since the method is based on mark allocation to assessment objectives it can be adopted for any subject offered by the

Zimbabwe School Examinations Council. In that regard, the study makes a distinct contribution to knowledge for the stakeholders of examinations in Zimbabwe.

Question Papers' Content Validity

It has been argued in earlier chapters that one of the instruments that is used to set and maintain standards of an examination is a question paper. This is the instrument that is used by examination boards, the world over, to assess the competencies that students should demonstrate to have achieved at various levels of education. The evidence on how the 1996, 1997 and 1998 question papers related to the expected standards was presented in Chapter IV.

It was shown in Chapter IV that there was reference to domains, assessment objectives and assessment schemes in syllabus documents when question papers were set in Geography 2248 and Integrated Science 5006. The forms that were completed by the experts mentioned in Chapter IV indicate that syllabus documents in the two subjects were the point of reference in the construction of the question papers. These syllabus documents are sent out to schools so that teachers and candidates have knowledge of the content and structure of the examinations, the assessment objectives upon which examinations would be set, and weighting of each assessment objective, etc. The fact that the syllabuses of the two subjects are there and are sent out to schools at the start of each school course shows that the Zimbabwe School Examinations Council is concerned with the maintenance of examination standards in the country.

This study established that the number of sections that are in the subject of Geography 2248 question papers was consistent over the three year period. The sections are Mapwork, Physical Environment, Economic Geography, and Population, Settlement and Trade. However, the Mapwork section was in Paper One only but consistently so over the three-year period. The consistency in the content that was tested in the two papers over the three years, as shown by the sections of the syllabuses mentioned above, was a measure of how standards were maintained from the point of view of the syllabus areas from where questions were set by the examining authority, the Zimbabwe School Examinations Council. If the content that was examined had varied from year to year then testing instruments would not have been valid. In relation to the second research question that was raised in Chapter I on content validity, evidence has been given to establish that there was high content validity in Geography 2248 over the three-year period.

The number of questions that were set in an examination is also a measure of consistency and must indicate to those concerned with the quality of examinations whether or not examination standards are maintained from one year to the next. Table 21 in Chapter IV shows the number of questions in each question paper. In Geography 2248 Paper One, the number of questions given to candidates each year was forty. The standard was maintained. Though the number of questions in Paper Two was nine in the three examinations under study there was a variation in the number of sub - questions in 1998. There was a close relationship between the number of questions and the content from which questions came from (see Tables 12 and 16). This is an

important aspect of comparing standards in that candidates are subjected to the same content areas from one year to the next. The number of questions and the content of the Geography question papers was a reflection of the demands of the syllabus document. What has been shown by this study is a desirable relationship between Geography question papers and the demands of the syllabus. The standard set by the syllabus was followed and maintained in the setting of Geography question papers. However, the number of sub-questions in Paper 2 varied. The number of sub-questions was 47 in 1998 while the two years, 1996 and 1997, had 50 each (see Table 21). This means that candidates who sat for the 1998 Geography Paper 2 examination were subjected to fewer questions than those of the other years. The number of sub-questions which caused the variations that was observed was very small. The implications on standards seem to have been neglible as shown by the consistency in the mean grades over the period. No better reason can be given for the variation except to point out that the syllabus document which guides question setters fails to be precise on the acceptable limits of the variations. However, it is recommended that this variation be eliminated so that candidates in different years are subjected to the same number of tasks in examinations of the same content.

There were forty questions which candidates were expected to answer in Science Paper One in each of the examinations over the three-year period. The number was, therefore, consistent. This paper, like the Geography Paper One, was a multiple-choice paper. From the point of view of number of questions that constitute an examination, these two papers were consistent in maintaining that standard.

The lack of consistency in the number of questions from content areas over the period under study was a cause for worry. The reason for this lack of consistency must stem from the fact that the Science syllabus does not guide setters on the number of questions that must be set on each content area. This, it should be noticed, is unlike the Geography syllabus document that was discussed above. It has been shown in Chapter IV that in 1998 there were far more questions on **Agriculture** in Paper One than in any other year and in 1996 there were six more questions on **energy** than in any other year. The ideal situation is what was achieved in the content area of ***Science in the Community*** over the three-year period. The same number of questions was set in this area over the three-year period. The differences in the number of questions set on content areas were even bigger in Science Paper Three. It was only in 1997 when one question was set on ***Science in the Community*** in paper three while there were two questions on ***Science in Industry*** and one question on the same content area in the two other years. In 1998 there were two questions set on ***Energy*** while the other two years had one question each. This lack of consistency in the number of questions was marked in Integrated Science 5006.

The choice of content areas of the syllabus to include in Integrated Science question papers must be addressed. Where a question paper has distinct sections that are related to domains in the syllabus, it becomes easy to ensure that those domains have been included from year to year. Integrated Science papers did not indicate these sections. It is recommended that the question paper be split into sections that are related to domains

in the syllabus. An example is to state in the syllabus that questions in Integrated Science Papers will come from the syllabus areas as indicated below. Using the average of what is in Tables 14, 18 and 20 the number of questions that should be in each section was arrived at. The same can be done for Integrated Science Papers 2 and 3.

Table 37. Number Of Questions In Paper 1 From The Syllabus Areas

Content Area	**Number of Questions**
Agriculture	12
Industry	6
Energy	9
Structures	5
Community	8
Total	**40**

Table 38. Number Of Questions In Paper 2 From The Syllabus Areas

Content Area	**Number of Questions**
Agriculture	3
Industry	2
Energy	3
Structures	2
Community	2
Total	**12**

Table 39. Number Of Questions In Paper 3 From The Syllabus Areas

Content Area	**Number of Questions**
Agriculture	1
Industry	1
Energy	1
Structures	0
Community	1
Total	**4**

This recommended approach is important for two reasons. First, candidates will be aware of what syllabus topics to expect in question papers and so can easily focus on

the examination. Second, the approach provides an easy way of accounting for the quality of a question paper to the stakeholders.

Science question paper twos had different number of questions over the three-year period. The question paper had thirteen questions in 1996, ten each for 1997 and 1998. The number of all questions, including the sub questions, shows an inconsistency over the three-year period. Table 21 in Chapter IV shows that 1996, 1997 and 1998 had 46, 42 and 44 questions and sub-questions respectively. The difference in the number of questions is really not large but when the number of questions is consistent over a period of time, one can conclude that the examination would have a consistent measure of the tasks that candidates are supposed to undertake in it.

Science Paper Three showed (Table 21 Chapter IV) the greatest variation in the number of questions that candidates were given to answer. There were 14 in 1996, 34 in 1997 and 19 in 1998.

The evidence that was collected on the content of the Science question papers had serious implications on comparability of examination standards. Candidates who wrote these question papers were in fact given different content areas from one year to the next. An examining authority should not allow such latitude in the setting of question papers. The rules that the setters have to follow when setting must be elaborate and strict on the content areas to be used when setting question papers. There can be a

tendency of varying the assessment objectives which must be in the question papers where the number of questions also varies.

Although Science and Geography Paper Ones had over the three-year period the same number of questions this was not the same with the other papers. It can be concluded, therefore, that candidates in one year were not subjected to the same number of tasks as in another examination. It can be recalled from Chapter IV that some question papers differed with as many as twenty tasks from one year to the next. It is prudent to ask the question: ***what can stop one candidate who has answered more questions in an examination in one year to say that his/her grade C in Science in 1996 is better than a grade B in the 1998 examination where there is a variation in the number of questions from one year to the next?*** It is, therefore, recommended that the same number of questions and sub questions that constitute a question paper be the same all the time an examination is set. This can only be changed if the syllabus that governs the content and examination format of the examination has been changed.

The point that the question papers in Geography 2248 were valid instruments has been endorsed. However, content areas in the Science question papers were not strictly the same from one year to the next.

This study did not go on to establish whether or not the academic demands the 47 Geography questions presented to candidates in 1998 were the same as those in the 50 questions of 1996 and 1997. If it can be verified that the academic demands of the

questions, though different in number, were the same then a similar argument could also be presented. The argument is, why is it not possible to vary the number of questions from nine in Geography Paper Two to some other number from year to year? The point is that question papers that have a variation of the number of questions while at the same time maintaining the same academic demands are difficult to design. Variations in the number of questions, be they sub-questions or the main questions, is not recommended. That practice must be discontinued because it is quite possible in such circumstances to have unstable standards from one year to the next. The syllabus documents of the subjects that have been discussed in this study are weak in this area. It is recommended that syllabus documents be prescriptive in the number of sub questions that candidates are expected to answer. There is a real danger of candidates in different years being given different number of tasks in examinations. The question that must be answered would be when and who would judge that the variation in the number of sub-questions is too large when the syllabus does not state so. If the weighting of the assessment objectives is strictly adhered to, the number of sub-questions would be controlled because the number of marks that are allocated to each question would obviously limit the number of questions. However, this study has established that the syllabuses allow a variation in the allocation of marks to the assessment objectives. There is, therefore, a real danger here that standards can vary from year to year if this issue is not addressed.

This study is not advancing the argument that the number of questions in a question paper is the only thing that determines the level of difficulty of a question paper. One

needs also to consider the skills that the set questions demand from candidates. This is the next area of discussion.

Assessment Objectives

It can be recalled from Chapter I that one of the research questions of this study was to establish whether or not there was a match between the assessment objectives in the question papers and those outlined in the syllabuses. It can be concluded that question paper setters need to work harder to achieve what is demanded by the syllabuses. The question papers that were scrutinised showed no **strict** adherence to the demands of the syllabuses.

The skills that each question paper was supposed to assess were discussed in Chapter II. The proportion of questions directed to each of these skills was expressed as a percentage of the total marks that must be awarded for answering the question. Table 22 in Chapter IV shows that none of the question papers strictly adhered to the syllabus specifications in regard to the weighting of skills. The graphs clearly show the differences in the standards between what the syllabuses state and what was in question papers. The 1996 Geography Question Paper 1's skill weighting was way out of what the syllabus recommended. The judges referred to in Chapter IV agreed that the question paper went above the recommended weighting of 40% at Skill 2 by 17,5%. At Skill 1 it was lower than the recommended weighting by 13,5%. It can be noticed in Table 22 that only the 1997 question Paper 2 had skill weightings that were close to the syllabus specifications.

Lack of adherence to the weighting of skills is a major weakness of the instruments used here for measuring candidates' performance and thus adversely affected the comparability of standards. As pointed out in Chapter IV, the Geography syllabus points out that strict adherence to the skills weightings is in fact not a requirement when setting question papers. This is the root problem of failure to achieve comparable standards within a subject from one year to the next. The flexibility that was given to test developers by the syllabus must have been quantified so that the levels at which the variation become unacceptable are known. It is strongly recommended here that syllabus developers, who in this case are the Zimbabwe School Examinations Council and the Ministry of Education's department of Curriculum Development, quantify the variations that are acceptable. It has been pointed out in Chapter I that this department works closely with the Zimbabwe School Examinations Council. If there is **need** to give the setters some flexibility in the award of marks to different skills then an example would be that in each skill category the percentage of marks allocated should not be exceeded by more than a particular percentage. I have put emphasis on need because there could be a strong argument for each of the points, giving the setter flexibility and prescribing the strict allocation of marks to assessment objectives. Although the prescriptive approach is excellent for comparability of standards sometimes the precision that educationists are expected by society to show is impossible to achieve in the area of education. Whenever that number is exceeded then the setters of examination papers would be aware that they are compromising the comparability of standards from one year to the next. Adherence to the permitted weighting of the assessment objectives is the key to maintenance of comparable

standards. The weightings determine the marks that are awarded which in turn determines the number of questions to be set.

The study has also shown that Science Paper 1 and 2, except for the 1997 Science Paper 2, did not have questions in Skill 3 category. This skill is tested in Science paper 3. It can be recalled from earlier chapters that Science Paper 3 is called Alternative to the Practical component of that syllabus. It should, therefore, have questions with manipulative and evaluative skills.

The Integrated Science syllabus fails to give details of how question papers are composed. Although for Paper one the syllabus states that the paper consists of forty compulsory questions this is not enough for the test developer who sets the question papers, the teacher who prepares the students for the examination and the student who writes the examination. The information that is lacking has a bearing on maintenance of examination standards. The syllabus domains from where the questions would be set are not given; the number of questions that would come from each of the domains and the skills that candidates are expected to demonstrate on the questions are also not given for Paper 1. The matrix in the Geography syllabus that was discussed earlier, shows what the Science syllabus lacked. It is highly recommended that this matrix be developed for the Integrated Science question papers.

Integrated Science Paper 3 showed its consistency in testing the three skills though the weighting of the skills was not consistent. Skills 2 and 3 had high weightings. This

was commendable because the paper exposed candidates to practical skills that are demanded by the practical paper. However, it is recommended that the skills that are tested here be expressed using Bloom's scheme of taxonomy of educational objectives. This can make it easier for stakeholders who include parents, teachers, candidates, the examination authority, and the Curriculum Development Unit of the Ministry of Education, to know what is expected in testing instruments in as far as the recognisable behavioural objectives. It can be recalled from Chapter I that the assessment objectives for Integrated Science were presented as knowledge and understanding; handling of information, and experimental skills. The last two are all embracing. The way assessment objectives are presented makes one ask a number of questions. For example: Is there application of learnt skills in a situation when one is handling information? Do experimental skills show that candidates have evaluated a given problem? Do candidates make decisions during experiments?

If the answer to these questions is yes then syllabus developers would need to break down the 100% weighting of Paper Three into components in order to make it clear that a known number of questions with a known number of skills has been set from one year to the next. Although it has been argued earlier on that assessment objectives are hierarchical and competence in the high order skill shows that one has acquired the lower order skills, it would make it easier to account for all skills if questions in the two papers are set in the way Paper 1 and the more difficult Geography Paper 2 are set. It would be good for a credible examination system to have all the question papers set in a similar way so that the system is accountable to stakeholders. This study has shown

that the Integrated Science assessment scheme needs some improvement. Clear assessment objectives that have weightings can facilitate the monitoring of standards in examinations. It is recommended that the Science Paper 3 have a layout of assessment objectives such as the ones suggested below:

Experimental Skills can be broken down as

Knowledge and understanding

Application of learnt concepts

Evaluation

These skills would be weighted accordingly. It can be argued that in any experiment candidates would display knowledge with understanding before they demonstrate that they have acquired high order skills. Marks would be given to those who demonstrate knowledge only. Such candidates would obviously not pass the examination. Similarly there is no reason why Papers 1 and 2 should not have some questions with high order skills. It is recommended that all the papers be set in such a way as to stretch the minds of candidates. One needs to remember the point that is emphasized by Hambleton, Swaminathan and Rodgers (1991) on tests as instruments that measure ability. They pointed out that on one hand a score of zero tells us that the examinees' ability is low but provides no information about exactly how low and on the other hand when a candidate gets every question in a test correct, the score does not provide any information about exactly how high a candidate's ability is.

This point emphasizes the fact that tested skills in question papers must be spread from low order to high order so that candidates' competencies are judged more accurately.

The experimental skill category has in it the evaluative skills such as drawing up conclusions on information given, planning and organising experimental investigations and drawing up generalisations from experiments. This study has identified a weakness here in as far as what the syllabus states and what the question papers test. It is recommended that the scheme of assessment for this subject be revisited and the three skills be weighted and be included in all question papers.

It was shown in Chapter IV that the comparison of levels of difficulty of the question papers over the three-year period was not commendable when one compared the weightings in the syllabus and what was in the question papers. However, the standards set were pointing in the right direction. Improvement is needed in this area. It is recommended that statements of attainment be developed from these skills so that stakeholders would interpret candidates' performance in terms of what students know and are able to do. Having established that the testing instruments are based on assessment objectives in the syllabus, it is recommended that statements reflect performance of candidates in bands of marks as allocated to the said assessment objectives. The bands would indicate minimum competency and the most that candidates can demonstrate to have acquired in an examination. It is not possible to develop that interpretation of grades without suggesting grade cut- off points that are based on the performance of candidates on the assessment objectives. Our next area of discussion is to provide that framework of interpreting the scores.

Linking Assessment Objectives to Candidates Performance

The figures (assessment objective weightings) that are used in this discussion have been taken from the schemes of assessment in the syllabuses. The same can be done for any syllabuses at the GCE Ordinary level because they all have weightings of assessment objectives. The suggested marks have been calculated using the percentage of marks allocated to each assessment objective. The basis of the calculations was that when one scores half the marks that are allocated to an assessment objective he/she would have achieved the minimum competency on that objective.

The allocation of marks to assessment objectives in Geography 2248 as shown in Table 40 means that one needs to score the half marks as follows: 17.5%, 20% and 12.5% to make a total of 50% for the syllabus mark. There is nothing magical about the 50% score as the one that can be taken to represent a passing score. It has been used in this study because it represents the halfway mark where marks are out of one hundred. The marks allocated to the assessment objectives can be used in any way that is acceptable to come up with agreed upon cut off scores. The cut off scores that I recommend here represent a technique of coming up with an interpretation of the grades that are awarded at the GCE ordinary level. It is proposed that a candidate, who then achieves such a score would have achieved a minimum standard at "O" level. Once a 50% cut – off mark is accepted as the one that represents the minimum acceptable standard of performance one needs not to have achieved the marks in the percentage breakdown of 17.5, 20 and 12.5 but may accumulate marks from the first two assessment objectives only. It would not be possible for a candidate to score marks at the highest order skill without scoring on these two skills because the skills are hierarchical. If that happens it

means that the question setters would not have stuck to weightings of the three assessment objectives. A candidate could score 25% on knowledge with understanding and 25% on application of geographical skills and achieve the 50% of the syllabus marks. This should be perfectly alright because the 50% is only reflecting the minimum competency level that is expected.

It is my recommendation that a candidate who achieves marks in the judgement and decision making assessment objective category should get at least grade B. Assessment objectives were described in Chapter II as hierarchical and that point supports the idea being discussed here. A 50% score would be a grade C while grade B should be at 76 - 87% of the marks. This is because a candidate would have to score all the marks in the knowledge with understanding and the application of geographical skills bands to be of grade B category to give a total of 75. Such a candidate should also score some points in the judgement and decision making category. The grade B band would go up to 87% while grade A would start at 88% to 100%. Notice that the calculations have been based on the requirement that in order to pass at grade A, a candidate must score all the marks at the first two levels of assessment objectives as well as more than half of the marks at the highest order assessment objective of judgement and decision making. Grade D would be pegged at a quarter of the marks allocated to each assessment objective while grade E would be one eighth of the marks allocated to each assessment objective. Ungraded work would be lower than one eighth of the marks allocated to each of the assessment objectives. Table 40 below shows how the cut – off points at C, D, and E were arrived at.

Table 40. Cut-Off Scores In Percentages At Grades C, D And E

ASSESSMENT OBJECTIVE	WEIGHTING	½ OF THE WEIGHTING	¼ OF WEIGHTING	1/8 OF THE WEIGHTING
Knowledge with Understanding	35	17.5	8.75	4.375
Application of Skills	40	20	10	5
Judgement & Decision-Making	25	12.5	6.25	3.175
TOTALS	100	50	25	12.5
Grade		**C**	**D**	**E**

Table 40 must be looked at in conjunction with Table 41 below.

The reason for using this approach is that the standards have been seen to be relatively stable over the three-year period. The stability has been evidenced by the content of the question papers, the skills and the number of candidates that achieve a particular grade over a period of three years.

Table 41. Linking Scores And Grades To Performance

RANGE OF MARKS	GRADE Stanine	DESCRIPTION OF GRADE
0 – 23 (0 – 33)	U (6-9)	Candidates do not have the basic knowledge and understanding of geographical facts.
24 – 34 (37 – 39)	E (5)	Candidates demonstrate that they have a knowledge of geographical facts.
35 – 54 (40 – 44)	D (4)	Candidates show an understanding of geographical facts.
55 – 74 (45 – 49)	C (3)	Candidates can apply geographical skills in a given situation.
75 – 86 (50 – 54)	B (2)	Candidates competently demonstrate the application of geographical skills and can make some decisions on geographical situations.
87 – 100 (55 – 100)	A (1)	Excellent performance by candidates who can evaluate geographical phenomenon.

It was argued in Chapter II that the use of norm-referencing and criterion-referencing is viewed as complementary in standards setting. It was, therefore, important to subject the data collected to an interpretation that is normative. As described earlier the stanine method is a technique of standardising scores on a scale that has nine units. The study used this technique because currently the ZIMSEC "O" level examination results are also reported in nine numeric grades. The actual calculations that were used by ZIMSEC were not released to this researcher because it was considered confidential but literature (Cohen, Swerdlik & Phillips, 1996) discusses how stanine cut-off scores are arrived at. There is no criterion that is used when determining the cut-off score but the

mean grade achieved and the standard deviation. This means that when the mean and the standard deviation changes so would the cut-off scores.

The marks achieved by the 1996 Geography 2248 candidates were used. The mean mark was 35 and the standard deviation was 9. The calculations start with the cut – off scores at stanine 5. This means that for the 5th stanine a quarter of the standard deviation (9) was added to the mean (35) to get the cut-off score of 37. For the cut-off score for the 6th stanine a quarter of the standard deviation is subtracted from the mean to get 33. The other cut-off scores are calculated as follows:

1st stanine = ½ SD + stanine 2 cut-off score = 55

2nd Stanine = ½ SD + stanine 3 cut-off score = 50

3rd stanine = ½ SD + stanine 4 cut-off score = 45

4th Stanine = ½ SD + stanine 5 cut-off score = 40

5th Stanine = Mean + ¼ SD = 37

6th Stanine = Mean - ¼ SD = 33

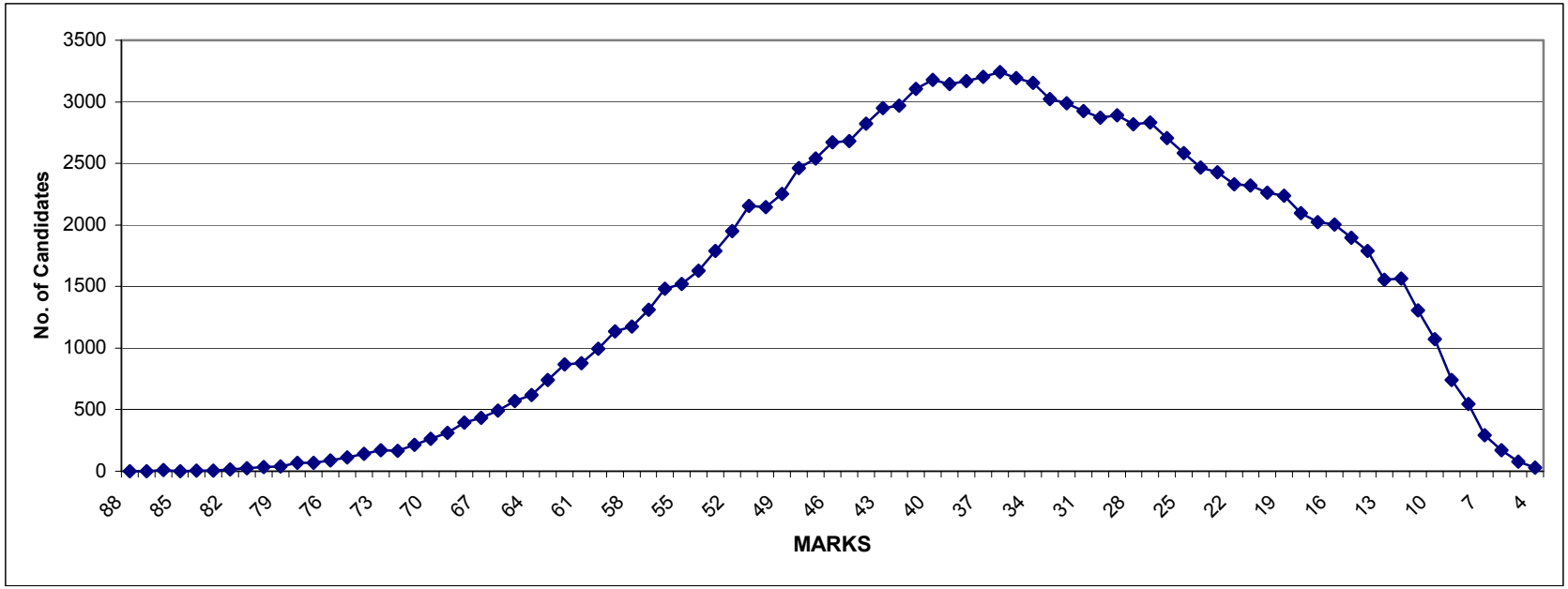
7th Stanine = Stanine 6 cut-off score - ½ SD = 29

8th Stanine = Stanine 7 cut-off score - ½ SD = 25

9th Stanine = Stanine cut-off score 8 - ½ SD = 21

Figure 9 shows the results of the calculations.

FIGURE 9. PERFORMANCE OF GEOGRAPHY CANDIDATES



The cut off scores would then be as follows:

Stanine	1	2	3	4	5	6	7	8	9
Cut–off Score	55	50	45	40	37	33	29	25	21

When all the cut-off scores that were yielded after the stanine system were compared to the ones already discussed, one finds out that there are some differences in the cut-off scores obtained after using the percentages of assessment objectives. The percentage of marks expected at each assessment objective is shown in Table 40. These are in the column headed weighting. At stanine 2 a candidate would have to score 50% of the marks meaning that all the marks for the knowledge with understanding assessment objective and only 15% of the marks for the application of skill assessment objective would have to be scored by candidates. There is only a difference of five marks between cut-off scores of stanine 1 and 2 and the percentages of marks at these stanines are far below those of the three assessment objectives. This is the shortcoming of this approach to determining standards. The candidates who are graded as the best would not have scored any marks on questions that are based on the highest order skill. Candidates would have to score some marks on questions that test the highest assessment objective to be deemed grade A in any subject. A comparison of the cut-off scores can be seen on Table 41. It would not be possible for the two techniques to be complimentary in this case.

The approach being presented by this study provides a working system to the determination of cut-off scores because the technique does not depend on the calibre of candidates who write an examination but on the weightings of the criteria.

I have argued that it is possible for the Zimbabwe School Examinations Council to state what each grade in each subject means in as far as the skills that are tested in question papers. The study has highlighted the skills that are in the question papers and the extent that they relate to syllabus requirements of Geography 2248 and Integrated Science 5006. The marks awarded in each syllabus must be at the same scale as all the syllabuses at the Ordinary level so that the level of competency in each syllabus can be compared easily by the same grades. The grade descriptions which have been proposed by this study are subject specific. However, a more general one which can be used to channel students into vocational areas and many other desired areas can be an area for further research. It would be possible to work out a statement of competency that relates to the number of questions that a candidate answers correctly. An example would be that when a candidate correctly answers questions of lower order skills only, the statement would be that the student could carry out simple instructions. The Zimbabwe School Examinations Council is challenged here to develop levels of attainment that are linked to scores that students can get in any assessment which the Council carries out.

In Chapter II it was argued that a way of reporting performance that is judged against known criteria would be more helpful to students, parents and employers. This study has revealed that the basis to do that is there in our examination system. Marks that are

allocated to each assessment objective have been used to link the expected test scores and the assessment objective and then provide a meaning to grades awarded at the GCE Ordinary level. The fact that examination standards have been shown to be relatively stable over a period of three years suggests the formulation of such criteria could be on that basis. Where there has been a cause for concern, as in the Integrated Science 5006 content validity of question papers, more work would need to be done by the Zimbabwe School Examinations Council before the use of the descriptions of grades.

Achieved Grades

One of the research questions of this study referred to the relationship of the achieved grades by candidates. The mean and mode of these grades were presented in Table 23 in Chapter IV. The reader is reminded that the numeric grades in the Graded Candidate Lists were used to find the mean and mode grades. This means that if fewer candidates were achieving grades such as 1, 2, or 3 it was more difficult for the candidates to achieve better grades, implying that it was not easy to score the high marks that were translated into grades. If the mean grades in Table 23 in Chapter IV were rounded to whole numbers this would show that except for the 1998 Geography Paper 1 with a mean grade of 5, all the papers and syllabuses had a mean grade of 6. Grades 5 and 6 correspond to a C-grade in alpha. The performance of candidates in Geography over the three-year period was, therefore, consistent. This is supported by the quality of the question papers that has already been discussed.

It is noted that candidates achieved better grades in Geography Paper 1 than in Paper 2 except for the 1996 examination when it was vice versa. The multiple-choice technique of assessment was used in Geography Paper 1 whereas in Paper 2 the structured essay technique was used. The results of the analysis of the grades leave us with the conclusion that multiple-choice type of questions are easier for candidates than structured essay type questions. This, therefore, means that setters at the Zimbabwe School Examinations Council were failing to pitch the questions at the same level as the structured essay type questions. Putting the argument differently I would say that the structured essay type questions were more difficult for the candidates than the multiple-choice questions. The Council needs to address the discrepancy in candidates' performance in the two techniques of answering questions. The smaller the numeric grade the easier it was to achieve a better grade. In this case Paper 1 was easier than Paper 2 in 1997 and 1998.

In Integrated Science the mean grade for the syllabus was the same over the three-year period. A mean grade of 6 was achieved. It should be pointed out that this was the same as in the subject of Geography indicating comparable standards of performance of the candidates who wrote the examinations. It has been shown in Chapter IV that candidates in Science produced better results in Paper One, which is a multiple choice question paper. This performance was similar in Geography Paper One which is also a multiple-choice type. The mode grade for Science Paper 1 over the three-year period was 1. This means that this paper was the easiest of the three papers that are offered in Integrated Science. It can be concluded that candidates found the multiple-choice technique of

assessment easier than the techniques that was used in the other papers. Question setters should always strive to ensure that multiple choice questions discriminate candidates who have acquired a lot of knowledge from those who would not have. For instance, it has been shown here that Science Paper 1 does not offer candidates questions at Skill 3. Evidence was presented to show that the distribution of skills in the two subjects was not strictly adhered to and so the question papers carried fewer questions at Skill 3.

The mean grades over the three-year period showed that the standard of the performance of candidates was comparable. In Geography, the syllabus mean grades over the three-year period were 5.7, 6.1 and 6.2. It can be concluded that it was not higher or lower by more than one grade. In Science the syllabus mean grades were 5.6, 5.5 and 5.7. It can be concluded that the standard of performance was comparable in the years 1996, 1997 and 1998. However, the 1997 Geography Paper 2 was the most difficult paper. The mean grade was 6.4. This was the lowest amongst the papers. The problem of lack of consistency in the skills and content in Integrated Science papers cannot be seen in the syllabus mean grades calculated here. The problem has been camouflaged by the aggregation of marks. The high grades in Paper 1 compensated for the low grades achieved in Papers 2 and 3.

One indicator for comparability of standards within a subject over a period of time that was discussed in Chapter II was the similarity of the number of candidates that achieve particular grades. Using this criterion, it can be concluded that comparable standards in Geography were in 1996 and 1998, and in Science they were in 1996 and 1997. Tables

27 and 29 in chapter IV showed that the number of candidates who passed the examination in the two subjects (grades 1-6) was about the same in the years mentioned above. However, in 1997 the number of candidates achieving the passing grades in Geography went down from 48.4% in 1996 to 40.2% in 1997. In 1998 it was 50.8%. This showed that the examination was the most difficult in 1997. This performance is supported by the skills tested as shown in Table 22. However, in general terms, the standard from the point of view of the number of candidates achieving pass grades is comparable over the years except the years mentioned above. The Council must continue to ensure that the stages where standards are set are strictly monitored.

The percentage of candidates achieving grades 1-6 (passing grades) in the Science syllabus was 46.5, 48.7 and 51.8 for 1996,1997 and 1998 respectively. This shows a fairly consistent award of grades over the three year period in Science. This indicates that the relationship between the grades achieved by candidates over the three-year period was good. However, there was also evidence to show that it was easier for candidates to achieve better passes in Paper 1 than in Papers 2 and 3. Therefore, it is recommended that the Zimbabwe School Examinations Council produces multiple-choice and structured questions that are comparable so that there is parity in what the question papers measure.

The box and whisker plots shown in Chapter IV illustrate that there is a relationship between the achieved grades in each of the subjects and also between them over the three-year period. These plots show the relationship of the spread of grades. The

distributions show that the standards that were set by the question papers were comparable.

The study suggests that it was easier to achieve higher grades in Science than in Geography. It can, therefore, be concluded that Integrated Science was slightly easier than Geography over the three-year period. This conclusion is based on the achieved grades and was supported by the distribution of skills that were found by a team of judges that scrutinised the skills that were tested by the question papers in the two subjects. Whereas the skills in the Geography syllabus and the subject content that was in the question papers conformed to the requirements in the syllabus, the same could not be established in Science. Science paper ones registered a performance that was really out of step with the other two Science papers. Candidates achieved high grades in this paper.

It was proved that there was a relationship in the mean grades of candidates who wrote the subjects in each of the three years and the F ratios that were calculated showed that the sample means were the same as those of the populations from which the data came from although the 1996 sample means were slightly higher than the critical value at 0,01 significance level.

The analysis of the achieved grades by candidates who wrote the two subjects shows that there was a positive relationship between the grades awarded. Standards were, therefore, comparable.

Consistency in The Award of Grades

The correlation coefficients of the Geography grades showed an impressive consistency. Though the correlation coefficient for the 1996 and 1997 Geography syllabuses was 0.4, those of 1997 and 1998 as well as the 1996 and the 1998 syllabuses were consistent at 0.5. The correlation coefficient of the grades in Integrated Science 5006 also showed consistency at 0.5 for the 1996 and 1997 syllabuses, and also for the 1997 and 1998 syllabuses. However, the 1996 and 1998 examinations' correlation coefficient was 0.4. These correlation coefficients are a high indicator of a very good relationship between grades. It can be recalled that a table that was used to interpret correlation coefficients was discussed in Chapter III. These correlation coefficients mean that standards, in as far as the award of consistent grades within each subject and between them, were comparable. This means that the examination system was performing in a manner which stakeholders expect it, which was to show consistency from year to year and from subject to subject.

The comparison of performance of candidates in the two subjects showed interesting results. This means that the performance of candidates in the two subjects, as a measure of standards, was comparable and, therefore, examination standards cannot be said to have slipped in the three-year period under study.

It should be remembered that the judges found out that there were no questions in Science Paper 1 that had Skill 3. This has been confirmed by the high mean grades in

these papers. High mean grades are reflective of easy question papers. Candidates achieved better grades in these papers over the three-year period. Where judges found many Skill 3 questions such as in Science Paper 3 the performance of candidates in such papers was very poor. The mode grade in Science Paper 3 was 9 while in Paper 1 it was 1.

The study has indicated that the setting of standards that was referred to by Rowntree (1985) as facing a possible danger of being more and more assessment procedures rather than standard attainments is not the case in Zimbabwe. Attributes of candidates' performance that are assessment objectives in the syllabuses and the content can be used at any time to evaluate question papers that carry examination standards.

The quantitative approach used in this study was supported by the qualitative data obtained from the experts on the content and the assessment objectives in the 1996, 1997 and 1998 question papers. The experts' qualitative comments showed that the quality of the papers was generally the same with the 1998 Geography Paper One being the easiest of the paper ones over the three year period. The correlation coefficients showed the close relationship between the papers in Geography over the three-year period. The correlation coefficient was consistent at 0.4. The experts also indicated that the Science paper ones were the easiest of the three papers offered in Science while the 1998 Science Paper three was the hardest. Overall, however, the experts agreed that the standard of the examinations were comparable. It is important to emphasize the point by referring to Matthews (1985) who noted that if a large

group of candidates takes both subject x and y, it should, if standards in x and y are equivalent, attain the same average grade in both.

This point has been observed in this study where the average alpha grade in each of the subject over the three years was C.

Conclusions and Recommendations

In summary, the main conclusions fall into six areas that are highlighted below. When these areas are put together the total picture of the degree of similarity of standards in the two subjects in this study is seen. The study has given evidence on the stability of standards and it is on that basis that a recommendation of reporting grades using assessment objectives was given. The conclusions are in the paragraphs that follow.

Examination standards in Geography 2248 and Integrated Science 5006, were in general, comparable from one year to the next. The comparability of standards was qualified as general because the correlation coefficients of candidates' performance in question papers and subjects was in the acceptable limits but the study identified some variations in the skills and number of questions that were given to the candidates in 1996, 1997 and 1998;

The assessment objectives that are in the syllabuses were found in the question papers as well. However, there was a variation in the skills that were tested. This variation consequently led to a disparity in the quality of question papers which candidates were

given. Though on average the variations were not very large there is a potential danger of standards slipping at the point of setting of question papers. The syllabus document is very weak on this area. It should state in categorical terms that weighting of assessment objectives should be stuck to religiously. If the document should allow variations the setters need to be told the acceptable variation limit;

The grades that were awarded to the same candidates who wrote Geography and Science are related. The mean grades were found to be the same and the correlation coefficients were consistent. This indicates that there is a comparable correlation coefficient between the grades that are awarded to candidates who wrote the Zimbabwe General Certificate of Education Geography 2248 and Integrated Science 5006;

Judges who participated in the study found out that the Geography question papers were valid measuring instruments of the syllabus objectives and content. The variations that were found in the content that was tested in Science led to the conclusion that Science syllabus developers and question paper setters did not have the skills that the Geography syllabus developers and setters had. There needs to be a standard structure of all the "O" level syllabuses in relation to skills and content that will be tested;

It is possible to develop statements that describe candidates' performance at the GCE "O" level. This study has shown how this can be done;

No one single factor can be used to determine whether or not there was comparability of examination standards within or between the subjects over the three-year period. I have discussed four factors and made recommendations. These four factors were central to this study. When the four factors are put together evidence indicated that there was a higher comparability of standards in Geography than in Integrated Science. The latitude which the Integrated Science 5006 syllabus gives to setters is the cause of the lack of high comparability of the examination standards. So this study found out that the more prescriptive a syllabus is in the content and skills to be tested the higher the comparability of standards. The prescription must be in the area of adherence to the number of tasks that are set in each examination and the weighting of the assessment objectives in question papers. The prescription in the way in which question papers are set should not cause problems in the teaching of the subjects. The prescription would mean that candidates would get question papers as they are described in the specification grids;

Future studies in this area could attempt to investigate the relationship between grades awarded by an examination board and the quality of scripts in a subject or in subjects from one year to the next. Although the study has presented evidence that when students take two subjects that have same standards the mean grades that are achieved by the candidates would be the same it is recommended that some future studies be also carried on the relationship of standards in other subjects.

This study contributes to an improvement of reporting examination standards in Zimbabwe in that grades awarded to candidates can now be linked to the assessment objectives which form the standards of our examinations. This has not yet been done in Zimbabwe and if adopted abilities of candidates would be compared in terms of what a candidate actually knows. So assessment objectives and not just a bald grade can be used to report standards in examinations in Zimbabwe.

REFERENCES

Adams, N. A., & Phillips, E.J. (1988). *General Certificate of Secondary Education. Inter-Group Comparability Study. Mathematics.* London: HMSO.

Bardell, G. S., Forest, G. M., & Shoesmith, D. J. (1978). *Comparability in GCE. A Review of the Boards' Studies, 1964 -1977*. Manchester: JMB Publication

Best, J. W., & Kahn, J. V. (1993). *Research in Education.* Boston: Allyn & Bacon.

Block, J. H. (1978). Standards and Criteria: A Response. *Journal of Educational Measurement.* 15. 291-295.

Borg, W. R., & Gall, D. M. (1993). *Education Research: An Introduction*. New York: Longman.

Brearley, A. (1986). *Report on The Needs of a Computer System.* Unpublished Report for the Examinations Branch of the Ministry of Education.

Chouhan, D. J. (1983). *The Localisation of Examinations*. Unpublished Report for the Examinations Branch of the Ministry of Education.

Christie, T., & Forrest, G. M. (1981). *Defining Examinations Standards.* London: Macmillan.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement.* 30. 93 – 106.

Cohen, L., & Manion, L. (1992). *Research Methods in Education.* London: Routledge.

Cohen, R. J., Swerdlik, M. E., & Phillips, S. M. (1996). *Psychological Testing and Assessment. An Introduction to Tests and Measurement*. London: Longman.

Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgements And Statistical Approaches. In H.

Goldstein & T. Lewis (Eds.), *Assessment-problems, Development and Statistical Issues: A Volume of Expert Contributions.* (p57-84). Chichester: John Wiley and Sons.

Croker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

DES. (1989). *The Task Group on Assessment and Testing Report.* London: HMSO.

Elliott, G. E., & Dexter, T. E. (1996). *An Annual Review of UCLES Grading Standards at the GCE "A" level 1995*. Unpublished paper from the University of Cambridge Local Examinations Syndicate

Elliott, G. L., & Massey, A. J. (1994). *Comparability of Standards in UCLES IGCSE Foreign Language French (June 1994) and MEG French (June 1994)*. Unpublished paper from the University of Cambridge Local Examinations Syndicate.

Erickson, B. H., & Nosanchuk, T. A. (1988). *Understanding Data An introduction to exploratory and confirmatory data analysis for students in the social sciences.* Milton Keynes, England: Open University Press.

Forrest, G. M., & Vickerman, C. (1982). *Standards in GCE: Subject Pairs Comparisons, 1972-1980.* Manchester: JMB Publication.

Forest, G. M., & Shoesmith, D. J. (1985). *A Second Review of GCE Comparability in GCE Studies.* Manchester: JMB Publication.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational Research: An Introduction* (6th Ed.). New York: Longman.

Glass, G. V. (1978). Standards and Criteria *Journal of Educational Measurement.* 15.

237-261.

Gipps, C. (1990). The Debate over Standards and the Uses of Testing. *British Journal of Education.* 36 21-36.

Goldstein, H. (1983). Measuring Changes in educational Attainment Over Time: Problems and Possibilities. *Journal of educational Measurement* 20. 369-377.

Goldstein, H., & Heath, D. (2000). *Educational Standards*. London: Macmillan.

Goldstein, H., & Lewis, T. (Eds.). (1996). *Assessment: Problems, Developments and Statistical Issues. A volume of expert contributions.* New York: John Wiley & Sons.

Gronlund, N. E. (1985). *Stating Objectives foe Classroom Instruction.* New York: Macmillan.

Hall, G. (1989). *Records of Achievement: Issues and Practice.* London: Kogan Page.

Hambleton, R. K., Swaminathan, H., & Rodgers, H. J. (1991). *Fundamentals of Item Response Theory.* London: Sage.

Jones, M. J., & Lotwick, W. R. (1979). *Report on the Inter-board Cross moderation Exercise in Biology at the Ordinary level,* 1978 (Unpublished).

Kachale, C. M. (1983). *Localisation of O and A Level Examinations* Unpublished Report for the Examinations Branch of the Ministry of Education.

Kane, M. (1998). Choosing Between Examinee-Centred and Test-centred Standard-Setting Methods. *Educational Assessment Journal.* 5. 129 – 145.

Kelly, A. V. (1999). *The Curriculum. Theory and Practice.* London: Paul Chapman Publishing.

Kolen, M. J. (1999). Threats to Score Comparability With Applications to Performance

Assessments and Computerized Adaptive Tests. *Educational Assessment Journal.* 6. 73 – 96.

Kubiszyn, T., & Borich, G. (1990). *Educational Testing and Measurement*. Glenview, Illinois: Scott, Foresman & Company.

Leedy, P. D. (1980). *Practical Research Planning and Design*. Second edition. London: Macmillan.

Linn, R., & Gronlund, N. E. (1995). *Measurement and Assessment in Teaching.* New Jersey: Prentice-Hall.

Mager, R. F. (1962). *Preparing Instructional Objectives*. Palo Alto, California: Fearson.

Maraire, N. (1982). *Report on Discussions with the UK Examination Boards on Localisation of the Examination System in Zimbabwe.* Unpublished Report for the Examinations Branch of the Ministry of Education.

Masango, R. B., & Nembaware, L. (1991). *The implementation on national goals and objectives in Zimbabwe and replacement of the University of Cambridge Local Examinations Syndicate's International "O" level Syllabuses by local syllabuses.* Unpublished paper presented at the International Association in Educational Assessment conference in Nairobi, Kenya.

Massey, A. J. (1997). *The Feasibility of equating National Test Standards in science between Key Stages 2 and 3 and from Year to Year.* Educational Review. 49 1 – 19.

Massey, A. J., & Dexter, T.E. (1995). *Using coursework marks to monitor grading standards in four large entry Midland Examining Group GCSE science*

syllabuses examined for the first time in 1994. Unpublished paper from the University of Cambridge Local Examinations Syndicate.

Matthews, J. C. (1985). *Examinations A Commentary.* London: George Allen & Unwin.

Methuen, B.O., Chih-Fen, K., & Burstein, L. (1991). Application of a New IRT-Based Detection Technique. *Journal of Educational Measurement* 28. 1-22.

Ministry of Education. (1983-1995). *Secretary for education Annual Reports* Harare: Government Printer.

Mukhurazhizha, T., & Masango, R. B. (1985). *Report on a trip to the Far East.* Unpublished Report for the Examinations Branch of the Ministry of Education.

Murphy, R., & Torrance, H. (1988). *The Changing Face of Educational Assessment.* Milton Keynes, England: Open University Press.

Newbould, C.A., & Scanlon, L. A. J. (Undated Paper). *Comparability of Standards in Public Examinations: A study of Monitoring Bias.* Unpublished paper from the University of Cambridge Local Examinations Syndicate.

Nikto, A. J. (1996). *Educational Assessment of Students* (2nd Edition). New Jersey: Prentice-Hall.

Northern Examining Authority Research Advisory Sub-Committee (1988) *Subject Pairs Comparisons by Grade.* Unpublished paper from the University of Cambridge Local Examinations Syndicate.

Orlich, D. C. (1978). *Designing Sensible Surveys.* London: Redgrave Publishing Co.

Patrick, F. (1996). Comparability of Scores From Performance Assessments. *Educational Measurement: Issues and Practice, 14. 13-15.*

Popham, W. J. (1990). *Modern Educational Measurement: A Practitioners' Perspective*. New Jersey: Prentice-Hall

Research Advisory Committee for the GCE Examining Boards. (1990). *Standards in Advanced Level Mathematics*. Unpublished paper from the University of Cambridge Local Examinations Syndicate.

Riddell, A.R. & Nyagura, L. M. (1991). *What Causes Differences in Achievement in Zimbabwe's Secondary Schools?* World Bank Sponsored Working Paper.

Rowntree, D. (1981). *Statistics Without Tears.* London: Penguin Books.

Rowntree, D. (1987). *Assessing Students: How Shall We Know Them?* London: Kogan Page.

Satterly, D. (1981). *Assessment in Schools*. Oxford: Blackwell.

School Council. (1979). *Standards in Public Examinations: Problems and Possibilities.* London: School Council.

Scriven, M. (1978). How to Anchor Standards. *Journal of educational Measurement:* 15. 273 – 275.

Secretary of Education. (1985). *Annual Report.* Harare: Government Printer.

Sedon, G. M. (1987). A Method of Item Analysis and Item Selection for the Construction of Criterion Referenced Tests. *British Journal of Educational Psychology* 57. 371-379.

Senior Secondary Assessment Board of South Australia. (SSABSA, 1998). *The Validity of the 1998 Biology Examinations.* Unpublished paper from SSABSA.

Stark, R. (1999). Measuring Science Standards in Scottish Schools: the assessment of

achievement programme. *Assessment in Education:* 6. 75-87

Sykes, R. C., & Fitzpatrick, A. R. (1992). The Stability of IRT b Values. *Journal of Educational Measurement.* 29. 201-211.

Tanyongana, N. (1985). *Report on a trip to the United Kingdom.* Unpublished Report for the Examinations Branch of the Ministry of Education.

Tuckman, B. W. (1978). *Conducting Educational Research.* New York: Harcourt Brace Jovanovich, INC.

Van der Linden, W.J. (1981). A Latent Look at Pretest-Posttest Validation of Criterion Referenced Test Items. *Review of Educational Research.* 51.379-402.

Van der Linden, W.J. (1982) A latent Trait For determining Interjudge Inconsistency in the Angoff and Nedelsky Techniques of Standard Setting. *Journal of educational Measurement.* 19. 208-231.

Vengesayi, C.K. (1991). *A comparative Study Of The Grade Seven Pretest General Paper and the Grade Seven Examination 1990 General Paper.* A Dissertation for the Master of Educational Psychology Course.

Weirsma, W., & Jurs S.G. (1990). *Educational Measurement and Testing.* Boston: Allyn and Bacon.

Wild, F. (1984). *Four Options for the establishment of the Zimbabwe Examinations Council.* Unpublished Report for the Examinations Branch of the Ministry of Education.

William, D. (1996). Meanings and Consequences in Standard Setting. *Assessment in Education.* 3. 47-61.

Willmott, A. S. (1980). *Twelve Years of examinations Research: 1965 – 1977.* London: Schools Council.

Wood, R. (1987). *Measyrement and Assessment in Education and Psychology.* London: Falmer Press

Wood, R., & Skurnik, L. S. (1969). *Item Banking: A Method For Producing School-based Examinations and Nationally Comparable Grades*. Slough, Bucks: NFER.

Zimbabwe Government. (1994). *The Zimbabwe School Examinations Council Act.* Harare: Government Printer.

Zimbabwe Government. (1996). *The Education Act.* Harare: Government Printer.

Zimbabwe School Examinations Council. (1996). The Geography 2248 Syllabus

Zimbabwe School Examinations Council. (1996). The Integrated Science 5006 Syllabus

APPENDIX A: CAMBRIDGE GEOGRAPHY SYLLABUS OUTLINE

APPENDIX B: CAMBRIDGE SCIENCE SYLLABUS OUTLINE

APPENDIX C: SCIENCE 5006 SYLLABUS AIMS AND OBJECTIVES

APPENDIX D: GEOGRAPHY 2248 SYLLABUS AIMS AND OBJECTIVES

APPENDIX E: GRADE THRESHOLD RECOMMENDATION FORM

APPENDIX F: CLASSIFICATION OF SCHOOLS

APPENDIX G: CLASSIFICATION OF CONTENT AND SKILLS

APPENDIX H: SAMPLE OF THE GRADED CANDIDATE LIST

APPENDIX I: SAMPLE OF THE EXCEL OUTPUT OF NUMERIC GRADES

APPENDIX J: JUDGEMENTS ON SKILLS

APPENDIX K: PERMISSION TO PUBLISH QUESTION PAPERS

APPENDIX L: SAMPLE OF QUESTION PAPERS