Studying School-based Summative Assessments in High-stakes Examinations in Bhutan: A Question of Trust?

Dechen Dolkar

Towards fulfilment of the Master Degree in Educational Science and Technology

Educational Management, Evaluation and Assessment

Faculty of Behavioural Science

University of Twente

Under the Supervision of Dr. J. W. Luyten and co-mentored by Drs. M. Hendriks

June, 2009

Acknowledgements

Though one has the will and the ability, many a times one does not have the resources. I thank first of all, the Dutch tax payers who through the Netherlands Fellowship Programme made it possible for me to undertake this master programme at the University of Twente in the Netherlands. At a time when needs and wants are unlimited, your generosity as a people and as a country is greatly appreciated.

Poetry sustains, inspires, guides and motivates. For sustaining my mind and soul while away from home through the nourishment provided in the poems you shared, Professor George Halley, University of New Brunswick, Canada, I thank you.

Undertaking this research has been as humbling an experience as it has been educative, and acknowledgement is due in this regard first of all to my mentor. For helping me stay focused when distracted by many seemingly attractive methods of analysis, for expert and subtle guidance when confused, for the detailed feedback on work submitted and most of all, for your patience, thank you, Dr. J.W. Luyten. Drs. M.A Hendriks, track co-ordinator for Educational Management, Evaluation and Assessment; for your guidance in the initial stages of getting started, facilitating my study by arranging mentoring under Dr. Luyten, and for your feedback, I thank you.

The courses offered in the Educational Management, Evaluation and Assessment track provided the knowledge base for this study and helped widen along with my mind, the scope of this paper. I thank all my lecturers; Dr. A.J. Visscher, Prof. Dr. F.J.G. Janssens, Dr. H. J. Vos, Dr. J.P. Fox, Prof. Dr. J. Scheerens, Dr. J.W. Luyten, Dr. M.A. Hendriks and Dr. M. Meelisssen.

For the help rendered in my frantic search and hunt for relevant literature, I thank Dr. Doug McCurry, ACER, Australia.

The administration of the questionnaires in the 10 sample schools would not have been possible without the support and commitment of the teachers identified to support the study. Mr. Amber Rai, Mr. Dhiman, Mr. Imtiaz Ahmed, Mr. K.N. Sharma, Ms. Leki Wangmo, Ms. Sonam Zangmo, Mr. Tahalman Gajmere, Ms. Tenzin Wangmo, Mr. Til Bdr. Chhetri and Mr. Uday Mitra, thank you.

Life is much easier with a supportive environment at work. Dr. Phub Rinchen, Secretary, Bhutan Board of Examinations, Bhutan, for being a supportive head and for piloting my instruments; Chador Wangmo, for promptly compiling the data of the pilot; Mr. Sangay Tenzin, Controller of Examinations, for helping me with my office work; and Au Kesang for keeping me informed about developments at work, I thank you all.

My family, on whom I selfishly imposed a long year of separation, I hope one day I can make it worth your while as it has been for me. I thank you for your love and support, your faith in me and your sacrifice.

C. Abstract/Summary

Based on the model of institutional effects, this study explores the reliability of the school-based teacher assessments when it serves to supplement student achievement in high stakes examinations. With specific focus on performance of the Bhutan Certificate of Secondary Education (BCSE) 2008 graduates in English, the study reports on the reliability of the teacher-assessed, continuous assessment (CA) marks relative to its relationship with the marks scored in the BCSE examination, teachers' rating of student competencies and student-self rating. A quantitative, cross-sectional correlation study was followed utilizing a descriptive survey method. A survey was undertaken in 10 higher secondary schools in Bhutan, involving 26 class 10 English teachers and 365 BCSE 2008 graduates to establish measures of student competency in English listening and speaking skills through teachers' rating and student self-rating. Though results indicate moderate conformity among the measurements in the study within schools, results between schools indicate that schools where the students on average score high on BCSE exam tend to score relatively low CA averages and vice versa. That compared to the CA marks for student performance in English listening and speaking skills, the teachers' rating of students on the same skills relates and co-varies with the BCSE exam marks and the student-self rating to a higher degree. The results illustrating anomalies in the relationship and variance of the CA marks with the other measurements of the same construct and findings of higher agreement among the teachers' rating the students' rating and the examination marks than with the CA marks, suggest the probable influence of incentives as predicted by the model of institutional effects.

CONTENTS

1. The Introduction

The general trend towards decentralization is evident in education, where the decentralization of decision making at the school level has led to increased school autonomy. The decentralization movement in education is based, as explained in the introductory chapter of their study (Maslowski, Scheerens and Luyten, 2007), on reasons pertaining to; cost sharing, in keeping with the democratic principles of distribution of power, promoting ownership among local stakeholders and encouraging educational novelties to meet specific needs. They also note that decentralisation of decision making in education "… is also infused, or at least legitimated by the aspiration to enhance the quality of education" (p. 304).

Decentralization, evident also in the certification of qualifications by centralised examination systems, is based on financial motives of cost sharing, meeting student needs and improving the quality of the assessment. It also aims to reduce the significance associated with the central examination by incorporating school based assessments in the certification of the qualifications (Tam, 1977). Centralised examination systems incorporate a school-based internal assessment component towards the aggregate performance of a student in many countries, among others; in Australia, Bhutan, England, Hong Kong, India, The Netherlands and in Singapore.

Increased autonomy brought about by decentralization, however, also translates to increased accountability and quality assurance of schools by schools. The pressures of improving quality influenced by incentives offered schools through decentralized assessment components could lead to strategic behaviour and ultimately compromise the fairness of the criteria for selection into higher education and the labour market.

This study addresses the issue of trust in school based assessment of student achievement as a measure of student competencies when used by high-stakes examination systems, especially in consideration of the use made of the results in selection criteria for higher education and employment. Crooks (2004), states that the consequences of assessment, is an important consideration in the validity of the assessment of learning.

Given the high stakes of the BCSE (class 10) Examination not only for students but also for school and teachers, the study seeks to explore the influence of incentives offered through the decentralised assessment of the BCSE examination on school and teacher behaviour through the examination of the nature of the CA marks. Specifically, this study seeks to explore the reliability of school based continuous assessment (CA) marks as a measure of the BCSE graduates' listening and speaking competencies in English against established national indicators.

The paper starts with a contextual description in chapter 2 followed by an explanation of the research problem in context presented in chapter 3. Chapter 4 describing the exploration and derivation of the research questions based on the theoretical framework of the study, presents the research questions and the statement of the hypotheses. The review of literature, addressing issues related to the questions explored in the study is presented in chapter 5. The research design is described and illustrated in chapter 6. Chapter 7 focuses on the presentation of the research methodology including descriptions of the sample, instrumentation and measurements, data collection procedures and data analysis. The results

related to the research questions are presented in chapter 8. The discussion focusing on the evaluation and interpretations of the results is presented in chapter 9. The summary of the findings and their implications are presented in the conclusion in chapter 10. The recommendation based on the results is placed in chapter 11. The reference list and the appendices are placed under chapters 12 and 13 respectively.

2. Description of the context

Trochim (2005) compares writing a research paper to telling a story; that the story in the research project is based on specific findings. The description of the setting being imperative to both good story telling and reader comprehension of the development of the plot, the report begins with a description of the context. The story takes place in Bhutan. This short chapter provides a brief description of Bhutan, the country's history of educational development and growth, and the structure of its educational system.

2.1 Bhutan, the country

Bhutan, also called 'Druk Yul' meaning 'land of the peaceful dragon', is a small kingdom in the Himalayas with a land area of 38,394 square kilometres and a population figure of 646,851 in 2006 (Gross National Happiness Commission, 2009). Its location, almost always and effectively, is given with reference to China in the North and India in the South. Its elevation of about 160 metres in its southern plains to about 7500 metres in the mountainous north and its rich forests covering about three fourths of the total land area, have made Bhutan home to a diverse range of flora and fauna and one of the most sought after, yet elusive tourist destinations in the world.

The country is divided into twenty 'Dzongkhags' or districts for administrative purposes. The first national elections held on the 24th of March, 2008 marked the peaceful transition of the country from monarchy to a constitutional democracy. Necessitating a better educated polity to nurture and strengthen its institution, the advent of democracy provides the impetus for greater access to and improvement in the quality of education.

Though generation of hydroelectricity and tourism hold potential for further economic development, environment and culture preservation are equally important indicators in the planning and assessment of development in the country. Bhutan is best associated and recognised in the international community for its definition of the concept of Gross National Happiness (GNH) and the application of this unique scale in planning and measuring the country's progress. Emanating from discontentment with conventional measures of development based on Gross Domestic Product, GNH extends the vision of development beyond the mere achievement of means to focusing on ends. The concept of GNH places at the core of development efforts, the happiness and wellbeing of the people.

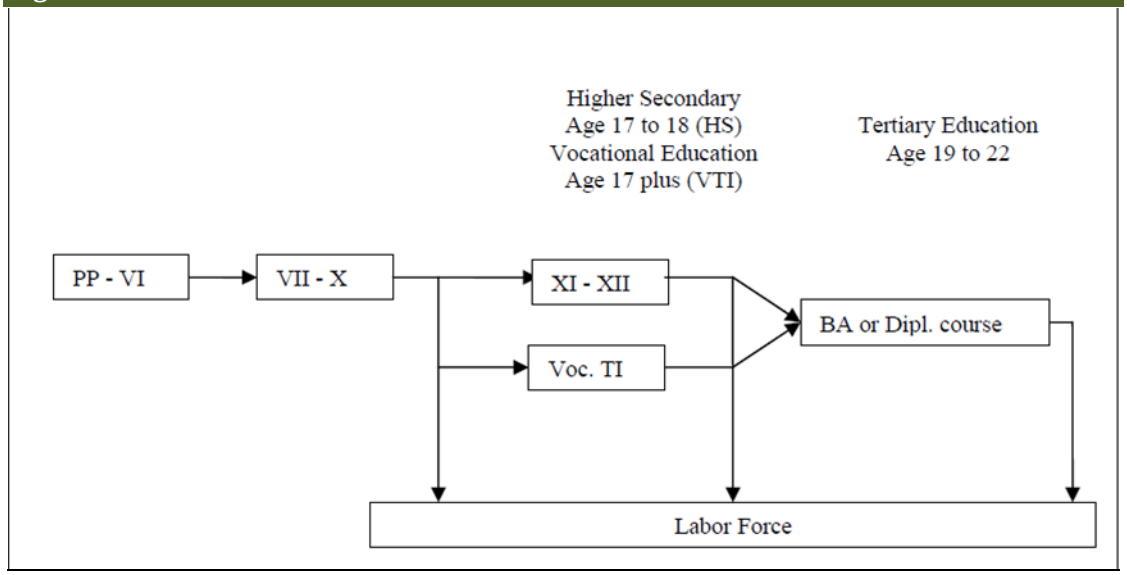
2.2 Education in Bhutan

Four decades ago, formal education in Bhutan was delivered through monastic education but after the introduction of Western Education in the early 1960s, Bhutan has made tremendous progress in the education sector. According to the General Statistics, Policy and Planning Division (PPD), Ministry of Education, Royal Government of Bhutan (RGoB), 2008, the total number of schools was 523; total student enrollment was approximately 157,112 and the number of teachers in all government and private schools including non-formal education instructors was 7,321 in 2008. Today the education system in Bhutan comprises general education, monastic education and non-formal education.

2.3 Structure of education in Bhutan

Formal education translates to 11 years of free and compulsory education. The 11 years comprise: 7 years in primary education from pre-primary to class 6; 2 years of lower secondary education from classes 7 to 8; and 2 years of middle secondary education from classes 9 to 10; beyond which, two years of higher secondary education is offered up to class 12. Schools may fall into either the lower secondary or middle secondary school category based on the level of education offered by the school.

Figure 1: General structure of education in Bhutan



(General Statistics, PPD, MoE, 2008, p.6)

3. Research problem in context

This chapter aims to provide an overview of the educational setting relevant to the study. The institutions and systems and the dynamics of their interaction which serve to describe the context informing the research is described in this chapter. Subsections 3.1 describes the centralized curriculum and national standards followed, 3.2 describes the centralized examination system, 3.3 explains the role of the English language in education in Bhutan, 3.4 explains the influence of English in the certification of qualifications and, 3.5 presents the immediate concerns informing this study.

3.1 Centralized Curriculum and National Standards

All schools in Bhutan, private and public, follow a centralized curriculum developed and prescribed by the Curriculum and Professional Support Service Division (CAPSD), Ministry of Education. The learning objectives specify the expected learning outcomes for every class and across the different subjects.

A framework of indicators against national standards has been specifically developed for English, 'The Silken Knot'. Standards in this document refer to statements of competencies that students should demonstrate after successful graduation from school in the five core skills in English study, viz. listening and speaking, writing, language and grammar, and reading and literature. Corresponding to standards in each skill are indicators of achievement specified across 8 levels. The indicators of achievement for the highest levels 7 and 8 correspond to the expected performance of students in classes 9 through to 12 (Centre for Educational Research and Development, 2002). The English curriculum is based on this framework. The learning objectives for the listening and speaking strands specified in the syllabus of all class levels are derived from the indicators of the national standards for English as documented in 'The Silken Knot'.

3.2 Centralized Examinations

The Bhutan Board of Examinations (BBE) conducts and certifies the following national level examinations: Bhutan Certificate of Secondary Education (BCSE) for class 10; and from 2006, Bhutan Higher Secondary Education Certificate (BHSEC) examination for class 12. The conduct and evaluation of the All Bhutan Class 6 (ABCLVI) examination for class 6 and the Lower Secondary School Certificate Examinations (LSSCE) for class 8, were decentralized at the school level in 1999 and 2006 respectively. The BBE continues to provide support to schools in the conduct of the class 6 examinations and serve monitoring functions by setting and providing schools the question papers, the marking scheme and the model answers for the examinations. The responsibility of setting standardized examination questions, marking schemes and model answers for the LSSCE has now been completely decentralized at the school level effective from 2009. This decision was passed at the 19th BBE Board Meeting, 8th May, 2009.

All schools are also affiliated to the Bhutan Board of Examinations (BBE) and hence it is mandatory for students all over the country to sit for these national examinations and successfully clear them to be able to gain admission to the next level of education. All subjects pursued by the students are examined through the external examinations conducted by the BBE.

The central examination incorporates both criterion-referenced and norm-referenced approaches. That the examinations are based on the assessment of learning outcomes, the questions reflect the expected standard of the class and that the results are based on student performance against the standards, make it criterion referenced. It is at the same time, norm-referenced since students' performance is reported as single scores for each subject allowing for rank ordering against the performance of other students.

3.2.1 The composition of the Bhutan Certificate of Secondary Examination (BCSE)

The Bhutan Certificate of Secondary Education (BCSE) comprises central examinations and a school based continuous assessment (CA) component in each subject. The mark awarded for performance in the central examination in each subject is scaled to 80%. Schools submit CA marks for each student over a total score of 20 which is aggregated with the marks attained in the BCSE examination. This aggregate mark is reflected as a single score for each subject in the results of the candidates.

Education is free and compulsory in Bhutan from pre-primary to class 10. Admission to all government owned higher education and vocational institutions is granted to those students who qualify by meeting the cut-off point set by the Department of School Education (DSE), Ministry of Education. The cut-off point is based on the availability of seats and the BCSE examination results. The BCSE examination also plays a filtering role in that students' overall performance and performance in different subjects in the examination determine the stream of study that a student is qualified to pursue. Higher secondary education in government schools is free and still remains the first option for parents and students alike. Students who do not qualify for admission to the government higher secondary schools may choose to continue in the private higher secondary schools that have been established. However, the majority, as well as the cream of the BCSE graduates is still to be found in government schools in Bhutan.

3.3 The Role of the English language in Education in Bhutan

English is the medium of instruction in Bhutan. All subjects, except the national language 'Dzongkha' are taught in English. The English language has been the medium of educational instruction since the beginning of modern education in the 1960s. Concerns have been raised over the past few years on the quality of Education with particular reference to the quality of English language competencies. The quality of education was brought up as a concern in the summer session of the country's National Assembly in May, 2006. Specific concerns have also been expressed with respect to the English speaking skills of high school graduates, who have reportedly failed to qualify for employment opportunities requiring such skills.

Measures to monitor the quality of education by the Ministry of Education have in fact, been undertaken as early as 2003, when at the primary level, a benchmark study was conducted for the first time in English literacy and numeracy in class 6 through the National Education Assessment (NEA), spearheaded by the Bhutan Board of Examinations (BBE). Benchmark studies for English literacy and numeracy were also undertaken for class 10 in November 2006. These studies have been instrumental in helping the Ministry of Education in ascertaining current benchmarks which will act as a basis to study the standards in future at regular intervals. The instruments used for testing English literacy in class 6 and

10 do not however, have a listening or speaking component and focuses on the other strands, i.e., writing, grammar, and reading and literature. Thus, as far as judging standards in English are concerned, listening and speaking is not included in the benchmark study. In attempting to ascertain and monitor the quality of English language competencies, the listening and speaking strands have yet to be addressed.

What makes the evaluation of the quality of English listening and speaking skills an elusive area is the mode of assessment for these two strands. Listening and speaking skills, not assessed formally in the BCSE examinations are assessed at the school level as continuous assessment (CA).

3.4 The Influence of English in the Certification of Qualifications

The Bhutan Certificate of Secondary Education (BCSE) is awarded upon successful completion of class 10. It comprises the year end BCSE examination which accounts for 80% of the total result for each subject, and a corresponding school-based continuous assessment (CA) component accounting for 20% of the total weighting of the results. The modes of assessment recommended for the school based internal assessment varies over the different subjects offered for the qualification.

English skills are assessed as two separate papers. The examination for paper I focuses on the assessment of Writing and Language and the school-based continuous assessment (CA) for paper I reports competency in Listening and Speaking skills. The written examination for English paper II assesses Reading and Literature and the assessment of student reading and writing portfolios informs the school-based continuous assessment.

The final award of marks for the BCSE at the end of class 10 for English Paper I comprises 80% written examination for Writing and Language conducted by BBE and 20% school-based continuous assessment (CA) for Listening and Speaking.

Table 1: Modes of Assessment for English Paper I for class 10

	Paper I	
Exam Component	Writing	Language
Exam weighting	50	30
CA component	Listening and Speaking	
CA weighting	20	
Total	100	

(CAPSD, MoE, 2005, p.142)

English is treated as one of the main subjects in which a student has to score a minimum of 40% to be awarded a pass certificate or certificate of successful completion.

3.5 Immediate concerns

Apart from a common framework of standards in the form of learning objectives and general subject guidelines for continuous assessment, central control mechanisms in the award of continuous

assessment (CA) marks by schools is limited. The CA marks reflect the assessment made by the independent teacher. Review with other teachers and committees are not suggested by the center.The standardization procedures to ensure comparability of marks are therefore, very basic and inadequate.

The support provided teachers for teaching and assessing listening and speaking skills in English in the form of a guide, the 'BCSE English Teacher's Guide' (CAPSD, 2005), has references to activities that may be undertaken in class to achieve the learning objectives. It has a one sentence statement on the use of viva voce within schools towards the assessment of students listening and speaking skills (p.110). Schools are prescribed the book 'Language Aloud…Allowed' for the teaching of listening and speaking. The forms in the book are *recommended* for record keeping of student performance (p. 111). The availability of this book in schools is an issue of concern. It may be safe to assume that, apart from the information that assessment of listening and speaking will contribute 20% towards performance in BCSE English paper I, there is not much support provided to teachers in the assessment of student listening and speaking competencies. In the absence of adequate standardization processes, there is concern regarding differences in interpretations of standards among teachers and schools in their assessment of student performance and award of CA marks.

There is no mechanism to monitor or moderate the CA marks reflecting students listening and speaking competency. Comparability of student performance in the CA marks is an issue that has yet to be addressed and its impact on the future of the students has yet to be studied. The issue of non-comparability between schools and want for studies on the nature of school based teachers' assessment in the form of CA marks, compromises its reliability as an indicator of the standard of listening and speaking competencies of the BCSE graduates.

In their study of the examination reform in Central and Eastern Europe, West and Crighton (1999) note the lack of comparability of student performance, which compromised fair treatment of students, as a reason that triggered assessment reforms. They state that non-comparability of student marks among schools, regions and over the years in otherwise tightly controlled systems, in times of stiff competition brought about by limited opportunities in life, was unfair to the students and did not serve its purpose of providing relevant and standardized feedback to the ministries and government to inform policy decisions. The findings of their study were based on an analysis of both local and international documents as well as on their personal experience.

Its publication, 'Rules and regulations for the conduct of public examinations in Bhutan' (2007) spells the motto of the Bhutan Board of Examinations:

- To conduct a fair assessment so that students get maximum opportunities to perform their best at the national level examinations and provide them with evaluation results that give a true picture of their performance.
- To support schools in the use of standardized testing system that guarantees a proper monitoring and a fair evaluation of the standard of achievement among pupils. (p. vii)

The context summary is marked by: the absence of central control mechanisms and need for adequate standardization mechanisms such as, a common framework of assessment criteria with detailed

descriptors of levels of achievement for each indicator in each class level; nonexistent moderating processes; and the resulting issues of non-comparability of student performance. While considering the importance of English and the high stakes of the BCSE certification, not only for students but also for the teachers and schools; there is concern as to whether the CA marks for listening and speaking in English, in keeping with the motto of 'fair assessment' of the Bhutan Board of Examinations, can be taken to inform the standard of level of student achievement.

This value of the motto of 'fair assessment' can be perceived in light of the use made of the results of the examinations in Bhutan and the impact it has on all the stakeholders as best described by Powdyel (2005):

> Many players have a stake in the examination results. They include: primarily, the students whose success is often measured in terms of the examination results; parents who pin their hopes and dreams on the success of their children; teachers whose performance is reflected in the achievements of their students; schools which construct their image and derive their strength from the performance of their pupils; the curriculum planners who want their intellectual architecture tested; and the government which provides the finance and wants good returns on its investment. (p. 50)

4 Exploration and Definition of the Research Questions

This chapter on the exploration and definition of the research questions begins with an explanation of the theoretical framework of the study in section 4.1. The theoretical problem is addressed in section 4.2. The application of the theoretical framework to the study is explained and illustrated in section 4.3, followed by the determination of the variables in 4.4. Section 4.5 is a statement of the research questions, and the hypotheses for the first question are stated in subsection 4.5.1.

4.1 Theoretical Framework

Empirical studies by Bishop based on data of the 1994-1995 Third International Maths and Science Study (TIMSS) on science and mathematics achievements of 13 year olds; the 1990-1991 International Association of the Evaluation of Educational Achievement's (IEA) Reading Study on reading literacy of 14 year olds; the1991 International Assessment of Educational Progress (IAEP) on science, maths and geography achievements of 13 year olds for 40, 24 and 15 countries respectively; have shown that centralised examinations serve to improve student performance (Bishop, 1998) (Bishop, 1999). Countries with Curriculum-based external exit examination systems outperformed those that did not.

The findings on the positive effects of centralized examinations on student performance have been corroborated by Fuchs and Woessmann in their study of international differences in student performance using data from Programme for International Student Assessment PISA. The study, based on the PISA 2000 data which included 32 countries focusing on educational achievements of 15 year olds in reading, maths and science, confirmed the institutional effect of centralized examinations through the observation of higher student performance in countries which had centralized examinations systems. In the study of international differences in educational institutions in 39 countries including 260,000 students, positive effects of centralized examinations on student performance were attributed to the examinations: encouraging the setting of higher standards; serving student and teacher monitoring purposes; serving to motivate both teachers and students; and encouraging better use of resources (Woessmann, 2000).

The study of the effects of test-based accountability on student achievement in the United States based on data from the National Assessment of Educational Progress (NAEP) State representative tests from 1992 onwards across 42 states, report that states with consequential accountability systems showed higher positive impacts on student performance. In exploring the operative mechanisms of accountability, the study refers to research on the motivational and strategic effects of external influences on shaping behaviour (Hanushek and Raymond, 2004).

The model of educational production approaches the school as an economic unit wherein production in terms of student achievement is an effect of the productivity of the workers within the school system. Bishop and Woessmann (2004) developed the theoretical model of institutional effects based on the application of the model of institutional economics to the school system. The model focuses on analysing the incentives that influence productivity of the workers and the institutional structures providing these incentives.

The model of institutional effects focuses on the relationship between only two actors of the schooling system, viz. the government and the student, where net benefits are sought by both actors in terms of their spending and effort respectively. On this basis, the theoretical model is extended to study the quality of educational production as being dependent on the institutional structures of the educational system. The model is used to study the effects of six main institutional features on the performance of students, among others; the institution of centralised examinations, the decentralization of decision making powers at the school level and teachers' influence. The theory proposes:

- the institution of centralised external examinations, by calling for transparency and accountability, is shown to provide an incentive to better student performance by; providing extrinsic motivation to learn, reducing the effects of peer pressure on learning, and by serving an effective monitoring function of both teachers and schools;
- behaviour of school and teacher as dependent on the nature of the incentives that are encouraged through decentralization of different areas of school system. That increased decentralization of decision making with respect to setting of level of standards and budget at the school and teacher level, provides incentives for opportunistic behaviour in the absence of constant monitoring mechanisms (Bishop and Woessmann, 2004).

4.2 Theoretical problem

The theoretical problem arises when aspects of the institution of centralized examination itself is decentralized at the school and teacher level, i.e., when standard setting is decentralised. Scheerens, Glas and Thomas (2007) state that when the results of external examinations have high stakes not only for students but also for the teachers and schools in its use for determining funding or school evaluations, strategic behaviour to enhance output is predictable. Quality assurance, generally reported by schools and evaluated by stakeholders, in terms of student performance and standard of achievement as a key indicator, increase the probability of the influence of incentives offered through the decentralised component of centralised examination systems.

Thus, incentives offered by the decentralization of reporting student achievement from the school level as a component of high stakes centralised examinations, make the reliability of school based Continuous Assessment (CA) marks as being indicative of, and reflecting students' level of achievement, questionable.

4.3 Application of the framework

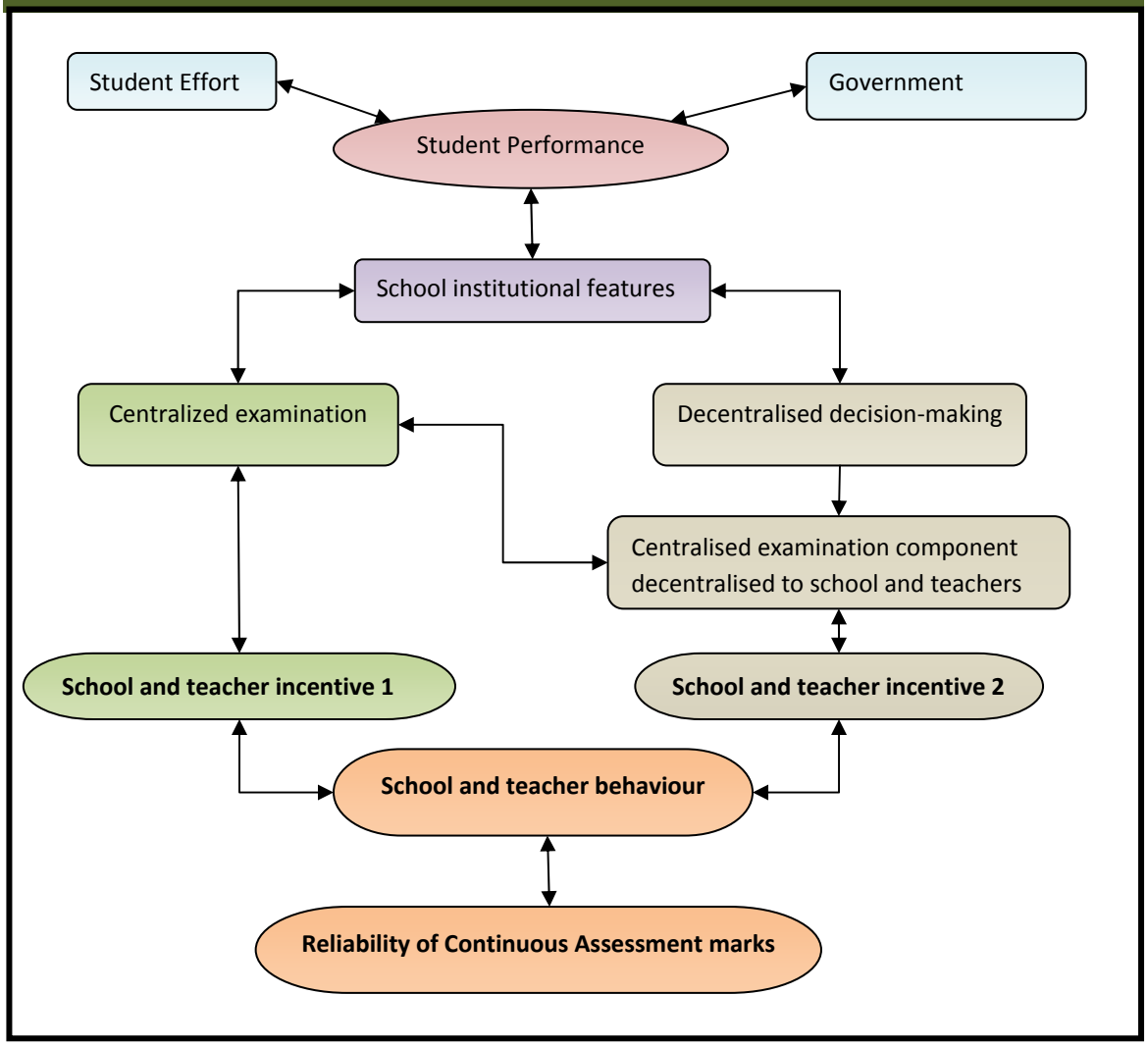
The study tested the model for the influence of incentives by studying school and teacher behaviour with respect to the reliability of the of the continuous assessment (CA) marks awarded by the teacher for student competency.

The incentives (*incentive 1 in model*) for school and teachers for improved student achievement are provided by the institution of centralised examinations. However, when components of the centralized examination is decentralised to the school and given the high stakes of the results of centralized examinations not only for students but also on the school and teachers, the influence of the incentives

(*incentive 2 in model*) may also interact. George Homans describes propositions in sociological theory as, "generalizations about how typical people act in typical situations" (Sanders and Pinhey, 1983, 17).

The model of institutional effects is extended as illustrated in figure 2, where benefits of government investment and student effort in dependent on the institutional structures, with the institution of centralised examinations providing incentives to better student performance. The proposition of increased decentralization of decision making with respect to setting of level of standards at the school and teacher level as providing incentives for opportunistic behaviour is tested in this study. The influence of the decentralised school-based assessment on school and teacher behaviour is investigated with respect to the reliability of the of the continuous assessment (CA) marks.

Figure 2: Extended model of institutional effects



"Reliability refers to the consistency with which a measure reflects a given performance level. A reliable measure should consistently reflect performance change when change occurs" (Drew, Hardman and Hosp, 2008). The reliability of the CA marks awarded by the teacher was explored relative to its

correlation with the examination marks for student achievement in the same subject. Triangulation of relationships with teachers' rating of student competency and student-self rating in the same study area were included for better exploration and explanation of relationships as well as to improve the validity of the findings.

Reliability in this study is thus, operationally defined as the association between the relative values of the four measurements.

4.4 The variables

In exploring the relationship of the CA marks with the BCSE exam marks, the study did not assume the BCSE standardised exam marks as the only benchmark against which to study the reliability of the CA marks. "The accuracy of summative teachers' assessment should be determined by the degree they reflect the learning goals and not through correlation with examinations results" (Harlen, 2008, pp. 223).

The first part of the study determined through a survey, measures of the BCSE 2008 graduates' English listening and speaking competencies against established indicators, in the form of teachers' ratings and student self-ratings.

The independent marks computed from the teachers' rating of student competency and the student-self ratings were used as variables to allow for more valid studies of relationships with the continuous assessment (CA) marks.

The reliability of the CA marks awarded by teachers for students' English listening and speaking competencies was explored relative to the following variables:

- BCSE exam marks of language and writing competencies
- Marks computed from teachers' ratings of student listening and speaking competencies
- Marks computed from student self-ratings of listening and speaking competencies

4.5 The research questions defined

In studying the reliability of the CA marks, its correlation with the BCSE exam marks was investigated relative to their relationship with the marks scored in teachers' rating and student-self ratings. The triangulation of relationships is expressed in the first research question:

What is the relationship between the extent to which goals for listening and speaking for BCSE 2008 graduates have been met as indicated by the CA marks, the teachers' ratings, the student self-rating and the extent to which goals for language and writing have been met as indicated by the BCSE exam marks?

To explore the influence of incentives, the study investigated the relationship of the CA marks and the marks scored in the three measurements (BCSE marks, teachers' rating and the student rating) relative

to the observations within the school and between the schools. The study also investigated for the influence of other factors by exploring the extent to which the CA marks could explain the variation in BCSE exam marks, relative to the extent to which teachers' and student-self ratings accounted for the variation.

What is the relationship of CA marks, the BCSE marks, teachers' rating and student rating within the school and across the schools and to what extent do the CA marks, relative to the other measurements, explain the variation in BCSE marks?

4.5.1 The statement of the hypotheses

The CA marks for student's listening and speaking competencies and teachers' and student self-rating are explicitly linked to the indicators of achievement in levels 7 and 8 of the standards for listening and speaking. Though the BCSE exam marks indicate student performance in language and writing and the CA marks for listening and speaking, both marks are aggregated to inform performance in English paper I and both measures are based on the same underlying dimension of English language skills. Within these considerations and in keeping with the findings from literature reviewed, hypotheses were formulated to address the first research question:

What is the relationship between the extent to which goals for listening and speaking for BCSE 2008 graduates have been met as indicated by the CA marks, the teachers' ratings, the student self-rating and the extent to which goals for language and writing have been met as indicated by the BCSE exam marks?

Studies on relationships between school-based teachers' assessment and standardised tests report correlations as low as 0.26 and as high as 0.92 (Hoge, 1989) (Kranjc, 2006). The study (Tam, 1977) focusing on moderation methods and procedures of incorporating internal assessments with the public examinations system in Hong Kong, presents different methods of moderation. Through the presentation of the moderation methods, the study notes the ineffectiveness of the entire moderation process when the correlation between internal assessments and reference tests are below + .3 (Elley and Livingstone as cited in Tam, 1977). A correlation of around + .7 is recommended between two forms of assessment for the meaningful application of moderation devices (Deale as cited in Tam, 1977).

The hypotheses on the size of the relationships between the measurements in the study are based on the benchmark of the correlation coefficient estimated for the CA marks (internal assessment) and BCSE marks (reference test) in hypothesis 3.

Hypothesis 1: CA marks will have strong positive correlation ≈ + .7 with teacher rating of student performance.

In determining the expected size of the correlation between CA marks and teachers' rating, the following facts were taken into consideration: that the teachers' rating and CA marks are student assessments of the same teacher; that both CA marks and teachers' rating reflect the assessment of student competency in the same language skills of listening and speaking competency and that the marks for both

measurements are assessments of the same students. Thus, the expectation of a relationship stronger than that between the CA marks and BCSE marks was inferred.

Hypothesis 2: CA marks will have a fair degree of positive correlation ≈ + .3 with student rating.

Guided by the literature on the relationship between teacher and student-self assessment, the factors affecting student-self assessment and in considering the novelty of the idea of student-self assessment in assessment practices in Bhutan, a moderate correlation was hypothesized between CA marks and student-self rating.

Hypothesis 3: CA marks will have strong positive correlation ≈ + .6 with BCSE exam marks.

The correlation coefficient in this hypothesis, guided by the applicability of moderation methods, was realistically estimated at r = +.6 to account for the different language skills being measured within the same language construct.

Hypothesis 4: There will be statistically significant positive relationships among the four measurements.

The continuous assessment (CA) marks, the teacher ratings, the student-self ratings and the BCSE exam marks measure student competency in English language skills. Thus, though the size of the relationships may differ, it was hypothesized that the four measurements be positively and significantly related to each other.

5. Review of Literature

Based on the research questions, the review of literature was guided by the underlying concern with regard to the school-based teachers' assessment and tests and examinations. Exploration and study of literature further highlighted the importance of reviewing the formative and summative nature of assessments and their differences with respect to nature and roles. A review of similar studies on the relationship between teachers' assessment and tests was undertaken to gain insights and guide the interpretation of the results. In the search for relevant literature, key terms such as *'correlation of scores' and 'internal or school based' and 'external or tests'* were employed. The University of Twente library search engine and the Google Scholar search using the library SFX were utilized to identify the database such as ERIC, JSTOR, Sage Journals and Scopus, to access the relevant journal articles. Continuously guided by the reference index of studies reviewed, a snowball approach was followed in the review of literature.

The chapter is a review of literature organised into two sections. Section 5.1 reviews school based teacher assessments, high-stakes centralised assessment and the implications of their amalgamation. The purpose of the review is to gain insights into the following: the nature and purpose of high stake centralized examinations (5.1.1); the nature and purpose of teachers' assessment (5.1.2); teachers' assessment in the certification of qualificarions (5.1.3); validity and reliability of teacher assessments for summative purposes in high stake examinations (5.1.4) and the relationship between teachers' assessment and standardized tests (5.1.5). A summary of the review is provided in the last subsection (5.1.6).

Since the study comprises a survey to collect teachers' rating of students and student self-rating, section 4.2, focuses on the review of literature on student self-assessment and teachers' assessment. The purpose of this review is to gain better understanding of the relationship between teachers' and student-self ratings and to investigate the validity and reliability of student-self rating. The review is organized under different subsections addressing the following concerns; the relationship between student-self assessment and teachers' assessment (5.2.1), factors influencing teachers' assessment (5.2.2), factors influencing student-self assessment (5.2.3), the validity and reliability of students self-assessment (5.2.4) and, teachers' and student assessments of oral performance (5.2.5). The last subsection presents a summary of the review (5.2.6).

5.1 Issues concerning the use of teacher assessments in high stake examinations

Studies on educational assessment invariably lead to a discussion on the summative and formative roles of assessment and tests, external examinations and continuous assessment where assessment is based on teachers' observation of students during class activity. It would serve the purpose of this study to touch upon these concepts to be able to better explore the intricacies surrounding centralized examinations and school based teachers' assessment and the resulting discomfort of one with the other.

5.1.1 The nature and purpose of high stake centralized examinations

It is important to consider the nature and purposes that are served by different forms of assessment before attempting to compare and correlate them. From this view, assessments can be seen as being either

summative or formative. The study of the forms of assessment makes a clear distinction between formative and summative assessment and between tests and school-based assessment. While tests fall under summative assessment, school-based teachers' assessment of student performance may be formative or summative depending on the purpose for which the assessment is used.

Studies of the literature in educational research on assessment (Kluger and Denisi, 1996), (Harlen and Crick, 2002) critique tests and examinations as tools for educational assessment, explaining the negative impact on learning and motivation to learn such as; encouraging learning through memorization and arresting creativity and innovation. The negative aspects of tests and examinations are highlighted as the merits of continuous teachers' assessment of student work.

However, despite the criticism of tests and examinations as tools of assessment for learning, they serve best, the specific purposes of the assessment of learning which is crucial in the educational setting and in the educational course of a student's life. Choi (1999) states that despite that negativity associated with examinations, they have yet to be abolished because it is widely accepted that some form of selection is unavoidable at certain stages, in either the endeavour for educational attainment or for employment, and because of the public confidence in examinations as a fair and objective selection mechanism.

Nagy, P. (2000), in his articles explaining the historical perspectives and the measurement theories behind the roles of assessment, identifies; gate-keeping, accountability, and instructional diagnosis, as the purposes served by external assessments.

Their paper (Kellaghan and Madaus, 2003) on external public examinations (as compiled in Harlen, 2008, volume 3) identify seven common characteristics of external examinations. External examinations share common characteristics as; institutions external to the school, have an overseeing function, are based on a prescribed curriculum, involve the administration of common tests at a given time, serve to certify qualifications, are voluntary in most contexts, and publish their results. In describing the role of examinations in education, Nuttall (as cited in Murphy and Broadfoot, 1995), in his study of the secondary school examination system in Britain highlights, 'the assessment of attainment, the maintenance of standards and licensing' as the functions of public examinations that serve selection procedures by institutes of higher education and employers.

In serving these purposes, it has been stated that centralised examinations also centralise curriculum control where countries performing well on international comparisons of educational standards: Germany, France, Japan and Singapore, have education systems that feature centralised curriculum and centralised examinations (Gipps, 1994).

In view of the purpose and useful function examinations serve in education and society, and in the absence of an alternative form of assessment which can serve, just as effectively, the specific functions highlighted, Nuttal (as cited in Murphy and Broadfoot, 1995), predicts that examinations will continue to serve its function well into the future.

5.1.2 The nature and purpose of teachers' assessment

The songs sung in favour of school-based teachers' assessment change tune when its summative role in the certification of high stake qualifications is investigated. The problem of school-based teacher assessments is the dilemma between the formative and summative purposes which they are intended to serve (Black and William, 1998).

School-based assessments can be formative or summative in nature and serving, therefore, two very different purposes. "In order for assessments to play its intended role, it is the purpose that ought to be the factor deciding the how, what, who, and when or the event" (Harlen, 2008, p. xxi). Based on Harlen's statement, the dual roles of school-based teachers' assessment may be seen to merge where the assessment of students' performance through ongoing assessment of activities and tasks by the teacher as part of the teaching learning process is then recorded to provide a summary at the end of a term as a measure of student performance. Thus in their role as contributors to the overall certification of qualifications, school-based teachers' assessment may employ formative assessment to inform and serve the summative purpose and thereby, differ in their composition and results from the examinations.

While agreeing with Maxwell's (2004) and Black's (2003) view of summative teacher assessments as facilitating learning and having a formative nature, it may be used for a variety of purposes; for internal purposes of record keeping and externally when it is used by examination bodies for certification of qualifications. In his review of research on the reliability and validity of teachers' summative assessments, Harlen (2005) states that the award of grades or marks to summarize learning counters the formative purpose of the assessment.

Thus, the nature of teachers' assessment may differ depending on the purpose it is intended to serve, i.e., either formative or summative, the precedence of one purpose resulting in the compromise of the other. Where teachers' assessment for learning assumes a more continuous form of feedback through interactions, validity of the assessment assumes priority over reliability. However, when it assumes a summative function, the equation is altered and dependent of the purpose of the summative assessment, reliability concerns vary.

5.1.3 Teachers' assessment in certification of qualifications

Though centralized examinations and school-based teachers' assessment of students differ with respect to their nature and focus of purpose, centralised examinations of countries such as Australia, Bhutan, Hong Kong, The Netherlands, Slovenia, Sweden and the UK include components of school-based teachers' assessment to broaden the range of learning outcomes assessed and to address the practicality of large scale assessment of certain learning outcomes. Countries in the African subcontinent such as Uganda (Pido, 2005) also incorporate teachers' assessment in central examinations.

Teachers' summative assessment is used, in the certification of qualifications through external examinations as a complementary component to cover important learning areas that are difficult to assess through examinations, as in the case of the General Certificate of Secondary Education (GCSE) in England and Wales (Harlen, 2005). The National Curriculum Assessment in England uses summative

teacher assessments' for measures of achievement in certain subject areas that are difficult to obtain at a national level through test, such as in listening and speaking for English, among others.

A common practise, in almost all high stake examination systems, is the monitoring and moderation of school-based teachers' assessment component. Aspects of the issue of trust in school-based teacher assessment is evident in the practices of examination boards as shown by the Hong Kong Examinations and Assessment Authority (HKEAA) in its application and use of School Based Assessment (SBA).

The HKEAA in its incorporation of the school based assessment (SBA) in the certification of the Hong Kong Certificate of Education Examination (HKCEE), documents that the SBA component in the certification of the HKCEE cannot and is not treated in the same manner as the examinations so that the SBA serves well the intended rationale for its inclusion. Its document explains that "…schools and teachers must be granted a certain degree of trust and autonomy in the design, implementation and specific timing of the assessment task." (2010 Hong Kong Certificate of Education Examination, p. 9). This autonomy and trust does not however, translate directly to trust in the marks submitted by the schools and teachers. School based assessment (SBA) marks sent by schools to the HKEAA are analysed, moderated for inter-teacher marking consistencies and adjusted to eliminate the marking inconsistencies. Moderation indices obtained from student performance in the examinations are used as references to compare the SBA marks and inform the necessary adjustments required.

Rowe, Turner and Lane, (as cited in Visscher and Coe, 2002) describe how the State of Victoria in Australia, in its certification of secondary schooling of Year 12 students through the Victorian Certificate of Education (VCE), include school-based summative teachers' assessment to enhance the range of learning outcomes assessed and towards improving the validity of the final assessment. Moderation methods changed overtime, from statistical moderation to external verification of sample student teacher-assessed work to large scale moderation against the General Achievement Test (GAT). The GAT scores serve to check for discrepancies in teacher assigned scores which may lead to expert review and finally to confirmation or change in the scores (Visscher and Coe, 2002). It is interesting to note that comprehensive moderating measures have had to be put into place to monitor school-based teachers' assessment. The use of the GAT, another test, against which to moderate school-based teacher's summative assessment is also an interesting measure despite the high construct validity teachers' summative assessment is accorded. Such measures appear to indicate the need for moderating influences to protect the reliability and consequential validity of the results.

The study of the case of Queensland by Withers (1987) and Butler (1995), where school-based teachers' assessment, having replaced external examinations in 1971, resulted in problems relating to quality and comparability of results at the state level as noted by Black and William (1998). Gipps and Stobart, in their study of the transition of assessment practices over thirty years, in the Australian State of Queensland where all examinations, except the Queensland Core Skills Test taken by seventeen year olds, was replaced by performance assessment by teachers; noted intensive moderation processes involving review of samples of student work by over 400 district review panels. The certification of the Queensland Core Skills Test which comprises a teacher assessment component is considered for admission to tertiary education (Gipps and Stobart as cited in Harlen, 2008). While lauding the transition of educational

assessment practices of the Australian State of Queensland as proof of the possibilities of using alternate teacher assessments in large scale examinations and to serve certification purposes, Gipps and Stobart (as cited in Harlen, 2008) state, "emphasis on accountability and selection is likely to deter such developments, as are legal challenges to the fairness of teachers' assessments" (p. 194).

Stiggins (1993) explains that the integration of centralised examinations and teacher assessments assume 'trickle-up' or 'trickle-down' patterns in which teacher assessment or centralized examinations inform each other respectively. The 'trickle up' or bottom-up approach refers to integrations of external examination and teacher assessment in systems where implementation of assessment is decentralised at the teacher level. Such assessments which serve teachers' needs by providing feedback on the teaching/learning are then aggregated to provide information to the centre. Stiggins states that such integration implies investment in teacher training, trust in teachers' assessment and the development of a management information system through which information can be communicated between schools and the next level of authority. The 'trickle down' or top-down approach refers to systems which have a common vision and set of outcomes decided at the central level, based on which teaching/learning objectives for class instruction and modes of assessment are prescribed at the school and teacher level. Such systems too, according to Stiggins, require investments in teacher training, training of independent test administrators and reliable information management systems to provide feedback of student performance to the schools. The relay of such results may or may not inform teachers and students needs. Thus, the limited scope and the costs of centralized assessments, possible disagreement on what constitutes student achievement affect the practicality of 'trickle-down' system while cost implications and issues of teacher assessment literacy and trust in teacher assessment affect 'trickle-up' approaches.

Based on his study of the composition of the Certificate of Secondary Education (CSE) in England which included 50% internal assessment comprising teacher assessment of students' oral work, project work and general course work, Nuttal (as cited in Murphy and Broadfoot, 1995) states that the system of school-based teachers' assessment of students requires a moderation processes so that there is uniformity in measurements across schools and is nationally valid.

5.1.4 Validity and reliability of teachers' summative assessments in high stake examinations

The University of Otago's policy on assessment states, "Validity is high for summative purposes when the assessment gives an accurate account of the student's capabilities at the time the final grade is awarded or the selection decision is made." (as cited by Crooks, 2004, ¶ 4). In explaining the contributions towards the validity of the assessment of learning, Crooks explains how the advantages and disadvantages of teachers' summative assessment translate as the disadvantages and advantages of examinations:

- teachers' summative assessment allows for assessment of important outcomes, facilitates multiple assessments and the observation and consideration of trends in performance, and reduces assessment stress;
- teachers' summative assessment has less control over the verification of ownership of work done; is less objective leading to the possible influence of factors other than the criteria for assessment and, teacher tendency to help students.

In the study on the reliability and validity of teachers' assessment used for summative purposes, a review of 30 out of a total of 431 existing literature using procedures and instruments of the Evidence for Policy and Practice Information and Co-ordinating (EPPI) Centre was undertaken by Harlen (2004). The EPPI is the Social Science research unit of the University of London. The study highlights the inverse relationship between reliability and validity and the degree to which the balance is altered depending on the purpose of the teachers' assessment. When the purpose of teachers' assessment is to serve a formative function, validity is seen to be more important than reliability and when the assessment assumes a summative purpose, reliability of assessment requires attention. The study reports that a complex balance between validity and reliability must be determined when teachers' assessment assumes a summative function. That the balance must be sought to ensure that the construct validity for which teacher assessment was preferred over tests is preserved while attempting to improve the reliability of the assessment. Detailed specification of criteria, teacher training in assessment to improve reliability of their ratings and moderating processes through professional collaboration are recommended to improve the validity and reliability of teachers' assessment (Harlen, 2004).

Stobart (2001) in his review of the validity of the national curriculum assessment in England, focusing on the assessments in key stage 2 and 3, notes that the teacher assessment component is devalued in the interpretation of the results with more focus being given to the test results. He states that undervaluing of teachers' assessments which comprises construct validity will lead to questions concerning the validity of the assessment system itself. However, in consideration of the purpose of the assessment, Nagi (2000) states, "Measurement issues with respect to accountability are more complex than those in the gatekeeping context. One central question is the trade-off between validity and reliability, or between curricular relevance and accuracy" (p. 268).

In their study exploring the community of assessment practice in the implementation of level descriptors in teachers' assessment (TA) in key stage 1, Hall and Harding (2002), studied, over a period of two years, six schools from six different local authorities (LEAs) in England. Using qualitative methods, structured interviews were conducted with all the year 2 and year 3 teachers, the school assessment coordinators and the LEA advisors. Besides reporting the lack of support for teachers in their interpretation and application of teachers' assessment and the concern among teachers of the lower value accorded to the teachers' assessment compared to the national test results, the study illustrates the professional dilemmas teachers face under the pressure of the publication of performance tables as commented by a teacher respondent:

> 'We've got to get the results up. And to do that legitimately is quite difficult. Up to now our attitude to TA (teacher assessment) has been "if you're not sure [about a higher level] don't give it" and now you're having to think "well, if you're not sure, we'd better give it". So you are really torn a lot about TA. You want to be honest with the children and with the Year 3 teacher. But outside agencies and the head are saying "get those levels up". What do you do? Do you, if you're not sure, give it? And then your Year 3 teacher is down on you like a ton of bricks. Or do you not give it and then have the head down on you like a ton of bricks? You're in a no-win situation.' (p. 12)

There is, therefore, indication that when the purpose of summative teachers' assessment has high stakes, such as informing the quality of the school through student achievement, incentives offered through the assessment at school level may influence school and teacher behaviour.

In addressing the integration of large scale centralised examinations and school-based classroom assessments, Scates (as cited by Stiggins, 1993) advises caution by illustrating clear distinctions between the two disciplines of educational assessment and highlighting differences in the background and training of teachers and testing agencies and the measurement methods employed. Stiggins summarizes the comparison of the two disciplines of assessment as differences in assumption where: assessors of standardized tests assume assessment to be matter of science and focus on technical quality and comparability of scores; whereas teacher assessment is more an interpersonal activity. That the one similarity of both approaches is the definition of the goals and roles of the assessor by their ethical and professional standards. Scates goes on to recommend keeping the two forms of assessment separate to ensure that they serve best the purpose for which they were intended and, allocating resources fairly between the two to ensure their quality. "Keep them separate, prepare teachers to assess well, let their assessment serve to promote ongoing student development, and randomly sample student performance periodically" (p. 104).

5.1.5 Relationship between teachers' assessment and standardized tests

There is very little literature published on recent studies undertaken to investigate the nature of school based teachers' summative assessment in relation to the central examinations in the certification of qualifications. Such studies appear to be conducted as action research by centralised examination systems specifically for in-house purposes. Some countries do share the procedures followed in the use of school based assessments, such as standardization and moderation processes, but very few studies specific to the questions raised in this study have been either addressed or published. A study in this area has been undertaken in The Netherlands (Lange and Dronkers, 2005) but is unfortunately published in Dutch as it is specifically intended to serve information needs of the society by the Dutch education system.

A review of literature (Hoge and Coladarci, 1989) to study the relationship between teachers' assessment against student performance in standardised tests involved studying and synthesizing 16 research studies. The studies reviewed differed in their methodological approach in terms of the link between teacher assessment and the standardised test; where most studies employed indirect estimations through teachers' ratings, while some used direct teacher estimations through teacher estimates of student performance in the same standardised test. Studies also differed in terms of the specificity of teacher assessment where some studies called for single score teachers' assessment while some employed item based teachers judgements. Despite these and other variations among the studies such as the unit of analysis used, the review reports correlations between teachers' judgement of student performance and standardised achievement tests as ranging between +.28 to +.92 with a median correlation of +.66.

In exploring the external and internal assessment of the matura, the certification of secondary education in Slovenia, Krajnc (2002) focusing on the subject Sociology, conducted a case study of 6 schools, 12 teachers and 17 % of the student population registered for the sociology examination. The matura comprises external examinations accounting for 75% of the final grade and the school based

seminar work accounting for the 25% towards the final grade. Applying a descriptive and causal non-experimental method, the study also comprised a survey of student ranking of experiences of seminar work. In its secondary analysis of the internal and external marks of the matura over nine years, the study reports low positive correlations. The correlations reported range between +.26 and +.60; with only year, 1999, reporting a correlation of +.60. A descriptive analysis of the survey reported more positive than negative student responses of experiences of seminar work.

In the empirical review of research comprising 30 studies on the validity of teachers' summative assessment using procedures and instruments of the EPPI- Centre, Harlen (2004 and 2005) concludes that summative teachers' assessment showed no correlation with standardized test scores for the same achievement in the absence of specific criteria. The study reports higher correlations between the two measurements with closer specification of tasks, use of assessment frameworks and finer specification of assessment criteria for teacher assessment. Harlen's review reports differences among schools and teachers in approaches to conducting teacher assessment. In investigating studies on teacher bias by student gender, special education needs and overall academic achievement, the review reports teacher training in assessment as influencing teacher judgement.

5.1.5.1 Difference in teachers' assessment across schools

Himmler and Schwager (2007), testing for the influence of student background on the lowering of educational standards through the inflation of decentralized grades of school, use data for 2002 and 2003 from the schooling system of the Netherlands in the certification of the MAVO, VBO HAVO and VWO. MAVO and VBO refer to the erstwhile streams of the middle level secondary education replaced, since 1999, by the VMBO. HAVO refers to five years of senior or higher general secondary education, and VWO refers to pre-university education for 6 years. Though there are two main streams in the form of general and vocational education, the Dutch education system allows for fluidity of student movement between vocational and general education at key levels of lower, middle and higher education. Difference in the standards established by the different schools was investigated through a comparative analysis of students' school grades and centralized examinations grades by taking the difference between the two scores which was then studied against variables estimated for social status. Though the study reports that schools with many students from disadvantaged backgrounds do tend to lower standards through inflation of the school assigned grades, it also cautions generalizations by noting that the PISA study on student achievement in Germany (Prenzel, et.al. 2005) reported different results of the central exams and the school grades being skewed against disadvantaged students even after having controlled for ability.

The study (Reeves, Boyle & Christie, 2001) based on data from the Qualifications and Curriculum Authority (QCA) School Sampling Project (SSP) in England exploress the relationship between student achievements in the National Curriculum tests, teachers' assessment and student characteristic variables. Though both test results and teacher assessment (TA) have equal status and though the TAs are reported alongside the test results, only the test results are employed for the publication of 'League Tables' of school performance. The TA is more for internal use by the schools to provide information to parents in addition to the test results. The study of the relationships focuses on Key Stage 2 standardised tests over a period of 3-year period (1996–98). Using the analysis of variance, school's identification was used as a school-level random factor to explain overall between-school differences in means. The findings of the

study report that while variables of student characteristics did not have any significant impacts to account for differences between the TAs and test results, schools as a factor was reported to have considerable impact on the size of the difference. Conducting a comparison of between school variations over the years, the study reports reduction in between school variance over the years though the overall relationship of the two measurements remains unchanged. The results of the study showing decrease in the differences between TA and test results among the schools over time, also implies positive correlation of means of measures between schools.

Willingham, Pollack, and Lewis (2002), note the paradox that while using school based grades to improve test validity and fairness, we do not trust grades to accurately measure educational outcomes. Willingham (2002) notes,

> Both grades and test scores play an important role in high-stakes educational decisions. Tests are often used because of uncertainty about the meaning of grades, yet grades are used to evaluate the validity and fairness of tests. Grades and tests provide this mutual support because it is commonly assumed that they do or should measure much the same thing. Yet the two measures often yield somewhat different results. (p.31)

In studying the variance between external test scores and school assigned grades, data from the National Education Longitudinal Study (NELS) of 1998 involving 8554 high school students from 581 schools was used to explore the role of the composition of five factors in explaining the variation. Rather than focusing on the structural differences between grades and test scores, the study (Willingham, et.al. 2002) analysed the role of the factors by focusing on trends in individual and group differences. Variation was studied in terms of differences in terms of student rank order of their high school grade average and NELS test score. In studying the relationship between the high school grade average and the NELS test score against teachers' ratings, the study reports teachers' ratings as having higher correlation with grades than with test scores and that grading variation among the schools accounted for most of the discrepancy between observed school grades and grades predicted from the test scores. Of the factors studied three considered to represent characteristic differences between test scores and grades were; grading variation, student characteristics and teachers' ratings. The results on the influence of grading variation as a factor reported that grading variation among schools was found to have a major influence accounting for differences between tests and grades.

Goldman and Hewitt (as cited in Willingham, et al., 2002) refer to the 'adaptive level' whereby teachers and schools have the tendency to adapt their grading standards to the ability of the students. The 'adaptive level' is reported as accounting for grading variation between different courses.

A study (Wikstrom and Wikstrom, 2005) similar in assessment practice to Bhutan was undertaken to study grade inflation and school competition with focus on the 1997 graduates from Swedish upper secondary schools. The decentralised grade setting at the school level and inadequate central control mechanisms, similarities shared with the decentralised assessment context in Bhutan informs the basis of the investigation. The grade point as well as the Swedish Scholastic Aptitude Test (SweSAT) has equal credence in the selection process for higher education. "Quality improvements following from competition may be limited by grade inflation and grade inflation may lead to mismatches in the selection

into higher education and in the labour market" (p. 310). The influence of the competition among schools brought about by the pressures of improving school quality and the incentives offered through the decentralised grading system on grade inflation are investigated.

The study (Wikstrom and Wikstrom, 2005) addresses grade inflation due to school competition relative to difference between schools (public and private) and between municipalities. The difference between the grade points and SweSAT scores of 22558 graduates of Swedish upper secondary schools were used to study grade inflation with background variables of student characteristics. The decentralised grade setting at the school level and inadequate central control mechanisms, similarities shared with the decentralised assessment context in Bhutan, informed the investigation.

The study (Wikstrom and Wikstrom, 2005) reports small effects with respect to grade inflation between schools and between municipalities with reference to potential competitive environments. The study presents strong evidence of grade inflation in private schools, where between the two schools types, score differences of male student is approximately 0.43 standard deviations and for females 0.24 standard deviations. "This means that an average 'male' student improves his position in the grade distribution by approximately 15% if graded in an independent school" (p. 317). Also reported is the difference is grade score by gender, where compared to males, females, on average, received higher grades by 0.5 standard deviations. Parent education related negatively to the difference in scores, is reported in the study as students with highly educated parents scoring higher on the SweSAT than in grades. This result may also be held to imply that relative to other students and to their own performance in the SweSAT, students with parents who were not highly educated received, on average, higher grades.

An earlier study (Thomas, Madaus, Raczek and Smees, 1998) based on the issue of equity in assessment, examined the relationship between student characteristics and student measures of achievement in three subject areas. The study employed the achievement scores of the 1992 summer term of 16,840 seven-year-old students in 538 schools in one large local education authority (LEA) in England. The student characteristics in the study include; gender, age, special education needs, family income and English as a second language. The subject areas studied were English, Mathematics and science. The study employed student achievement measured though scores attained in the National curriculum (NC) results which comprise two measurements, viz., teachers' assessment (TA) and standard task assessments (STs), both assessed by the teacher. For the reading component in English, the NC results were contrasted with a National Foundation for Educational Research (NFER) standardised, word recognition test.

The multilevel analysis of the study (Thomas et al., 1998), which examined among other, school level variation in teacher assessment and standard task assessments, reported that student background factors were more useful in explaining variation between schools for the (TA) in English than for other subjects. While background factors accounted for 47% of the variation between schools, schools explained just 11% of the variance in student scores. The explanation of this finding is attributed to the possibility that English is more likely to be influenced by background factors than other subjects since it can be learnt at home. In examining school level variation in student achievement, the results report 39% to 16% variation in student achievement in handwriting and spelling respectively, to be attributed by differences between schools. The difference between schools is explained as difference among teachers in the interpretation of the assessment criteria.

5.1.6 Summary of review

This subsection provides a summary of the literature reviewed on teachers' summative assessment and external examinations as might inform this study.

They being two different forms of assessment, school based teacher assessment and external examinations are intended to serve different purposes. Literature illustrates that the very dual purpose, both formative and summative that school based teachers' assessment is intended to serve, and the integration of both purposes, may explain why school based teachers' assessments may differ in composition from that of standardised tests and examinations.

The integration of teachers' assessment with high stake examinations to certify qualifications in different education systems differ with respect to the form and methodology of integration. Some examination systems report teachers' assessment a separate grades while it is aggregated to inform the total score in others. The process of integration can be centrally controlled and prescribed or schools may have more implementation control and provide information required by the centre. In both cases, the implications with respect to investment required in the form of information management systems teacher training and support is vital for successful and meaningful integration. Literature states the integration of teachers' assessment in high stakes examinations should focus on finding the right balance between upholding the inherent high construct validity of the teachers' assessment and the enhancing its reliability.

The literature on the incorporation of teachers' assessment in high-stakes examination systems highlights the importance of standardizations of the process as well as the moderation of outcomes of summative teacher assessments to ensure the consistency and reliability of the assessment by the institutes conducting and certifying centralized examinations. Though discrepancies observed are justified on the basis of the differences in nature and purpose of the assessments and as reflecting the strengths of their individual forms, reliability of assessment for purposes of gate-keeping and certification by centralised examinations indicate the need to correct observed differences between tests and grade scores through moderation procedures.

The standardization and moderation of teachers' assessment by examination systems that certify qualifications may be justified in consideration of the influence of the pressures schools and teachers face to improve the quality of the school through student achievement. Also taking into consideration the influence of the 'adaptive level' explained under 4.1.5.1.

Evidence of differences and variation in measurements of external examinations and teacher assessments is reported with both high and low correlations in different studies. Correlation coefficients reported between the two forms of assessment in the studies range from + .26 to + .92. The literature also suggests that correlation of teachers' summative assessment of students with examination scores alone is not a valid comparison in itself though measuring the same construct, due to the difference in nature of the assessment.

The difference in measurements and the variations are reported more across and between schools and much less within schools. The school itself is also reported as a factor accounting for differences

between the two forms of assessments with school characteristics, such as type of school explaining variation between the two measurements. Student characteristics, such as, ability, gender and background are determined as a factor influencing this variation.

5.2 Student self-assessment and teachers' assessment

Since the study involves determining a measure of student achievement in listening and speaking through student self-ratings and teachers' ratings, a review of literature on the subject is warranted for better guidance in the interpretation and reporting of its findings.

5.2.1 The relationship between student-self assessment and teachers' assessment

Elliot, DiPerna, Mroch and Lang, (2004) studied the relationship of teacher and student-self rating with focus on the differences in student attitudes and behaviours that influence the benefits of classroom instruction, referred to as 'academic enablers'. The study used teachers' ratings and student self-ratings and involved a national representative sample of teachers' rating of 2060 K-12 students and student ratings of 534 students from 80 schools across 30 states employing the Academic Competence Evaluation Scales (ACES) rating scale. The results showed medium to high correlations between the range of + .40 and + .60 between teacher and student self-ratings (Elliot, DiPerna, Mroch and Lang, 2004).

The meta-analysis of relevant studies undertaken to explain the relation of student-self assessment and teachers' assessment by Boud and Falchikov (1989a) mostly comprise studies which focus on students in higher professional and academic institutions with very few studies at the school level qualifying for the analysis. Considering, however, the scale of the study, its findings, may inform the interpretation of patterns associated in this field of study. By drawing together and comparing the results of all the studies included in the meta-analysis, Boud and Falchikov address questions on student-self assessment with reference to teachers' assessment. With regard to the question of students over-rating or under-rating, the findings report more cases of agreement than disagreement between student-self assessment and teachers' assessment. In cases showing disagreements, no consistency was found with respect to students' tendency to over-rate or under-rate their performance. In investigating rating patterns of students of different abilities, the review reports that weaker and less mature students tend to overestimate their achievement to a greater degree in comparison to the more able and mature students underestimations. Thus, abler students are better able to rate themselves in a manner consistent with their teacher ratings than low achievers. With respect to use of self rating for assessment purposes and student rating behavior, the findings suggest student tendency to overrate their achievements. Of the six studies that investigated the question of rating behavior and gender, only three showed differences in rating patterns between males and females. Though three studies (O'Neill, 1985; Hoffman and Geller, 1981; Jackson, 1988) suggested female-self ratings have better agreement with teachers' ratings and reported female tendency to underrate self achievement, the same studies reportedly showed variations in findings, with variation in context.

The follow up study by Boud and Falchikov (1989b), undertaken on the findings of the earlier study mentioned above, focused on indentifying factors affecting the correlation of student-self rating with teachers' ratings. The study involved a meta-analysis, whereby a common metric was established for

comparisons of the effects of the various studies being analyzed and established criteria for the selection of studies was used prior to statistical calculation of effects. The results of the study reported correlations between teachers' and student-self rating ranging as wide as + .05 and + .82. One marked observation of the analysis was the influence of the design of the study itself; studies with better designs reporting closer correlations between student and teachers' ratings. Also observed was the influence of the nature of the subject being assessed, with more accurate student-self assessment with respect to teacher assessments being reported in the study areas related to science.

5.2.2 Factors influencing teachers' assessment

Studies also report student characteristics such as; gender and behaviour, as factors influencing summative teachers' assessment and ratings in addition to student competence in the subject (Willingham, Pollack and Lewis, 2002) (Harlen, 2005).

5.2.3 Factors influencing student-self assessment

Gender as a variable influencing student self rating was investigated in a study involving 5440 students from grade 5 through 8 by Robitaille (1977). The study measured both students' feeling of self confidence to perform an arithmetic computation and their actual performance in the task to explore the relationship between self-confidence and achievement for girls and boys through multivariate analysis of variance. The findings show that despite better performance in the task, girls measured lower than boys for self confidence in given task; in other words, there is the tendency for boys to overrate and for girls to underrate their ability.

5.2.4 The validity and reliability of student self-assessment

In a more recent review of the study conducted by Boud and Falchikov, Ross (2006) investigates among others, questions concerning the reliability and validity of student-self assessments. The review is based on an investigative approach and relies heavily on the findings of existing literature to draw conclusions. Despite the apparent lack of any statistical confirmations, the study provides a good summary of studies undertaken earlier. On reliability of student assessment, the review reports better consistency for more mature students, for students trained in self assessment and for certain subject areas. The review of studies did not show any consistency in terms of informing the validity of student-self assessments where validity is equated against teacher rating; with different studies reporting contradictory findings.

5.2.5 Teachers' and student-self assessment of oral performance:

Teachers' assessment and student-self assessment in this study is specific to the assessment of listening and speaking skills. Chen (2008) states that in keeping with the theories of constructivism and autonomy of the learner, the teaching of language is incorporating student self-assessment to a greater degree. Studies by Blanche, 1988; Blue, 1994; Dickson, 1987 and Oskarsson, 1989 (as cited in Chen, 2008) recommend opportunities for students to assess their own levels of language competencies to provide for better focus on, and control of their own learning. Paradoxically, despite the benefits in terms

of learning and learner autonomy, Luoma and Tarnanen, 2003 (as cited in Chen, 2008) note that in practice, self assessment opportunities are rarely provided. This statement hold true in the case of Bhutan where, student-self assessment remains an unfamiliar terminology in assessment practices.

Chen (2008) investigated student learning of self-assessment in relation to teacher assessment of oral performance in English. The study involving 28 students enrolled for an oral training course in English in a national university in Southern Taiwan, involved training students in self and peer assessment as part of the training course. Using evaluation forms with criteria for assessment developed by teacher and students as well as questionnaires, data was collected at two points in time of the training course. The results of the correlation analysis between student and teachers' assessment report different indices for the first and second cycle; r = + .5521 and r = + .7938 respectively with p values for both less than 0.05. Chen concludes that the results confirm the contribution of practice to the accuracy of student-self assessments as reported in his previous study in 2006 and by AlFallay, 2004. In discussing the validity of student assessments of oral skills, studies (Patri, 2002) (Magin and Helmore, 2001) also make reference to their pedagogic values and positive influence on learning outcomes. This finding has implication on the use of student-self assessment in this study as student-self ratings were collected without any interventions in the form of training. The use of student self-assessment is not a feature reflected in the modes of assessment for listening and speaking, or any other skill or subject in Bhutan.

5.2.6 Summary of review

The literature suggests positive correlations between student self-assessment and teachers' assessments. The degree of positive correlations reported ranging as low as + 0.05 and as high as + 0.82, indicates the difficulty in drawing concrete conclusions as to the relationship of teachers' and student-self assessments. As literature purports, the range of the correlations appears to depend upon the following:

- nature of the subject being assessed, with higher correlations of the two assessments for science subjects.
- ability and maturity of students, with more abler and mature students showing assessments more consistent with the assessment of their teachers.
- design of the study, with studies with better design showing higher correlation between student assessment and teacher assessment.
- The familiarity of students with self assessment is also reported to influence relationship of student-self assessment with teachers' assessment, where encouraging and training students in self assessment of oral skills shows improved correlations with teachers' assessment.

Literature also suggests that teacher assessment is influenced by student gender, behaviour and competency in subject. Student self-assessment is found to be influenced by gender, with girls showing a tendency to underrate. Training and opportunities for self-assessment are shown to influence both learning and accuracy of self-assessment.

6. Design:

This chapter is devoted to presenting the design of the study. Section 6.1 states the assumptions on which the study is based. Section 6.2 explains and illustrates the design of the study and its subsection 6.2.1 discusses the internal validity of the design.

6.1 Assumptions of the study

In exploring the reliability of the continuous assessment (CA) marks, its relationship with the BCSE exam marks, the teachers' rating and the student-self rating were taken as the reference point.

Though validity and reliability of the BCSE exams are addressed in the study, the comparability of the BCSE exam marks must be noted as assumed to caution generalization of findings over time. The study being a one-time cross sectional measure, issues pertaining to the comparability of the BCSE exam marks have not been addressed.

6.2 Research design

Though the quantitative approach is positivistic and has been termed unimaginative, it is appropriate when seeking undeniable facts (Vaus, 2002). A quantitative, cross-sectional correlation study is made utilizing a descriptive survey method. "Correlation studies are used in research whenever we want to explore relationship or make predictions." (Charles, 1998: 10).

A descriptive survey method is recommended to allow for generalizations to be made from the sample to the population (Sanders and Pinhey, 1983**,** 127). The survey comprises teachers' ratings and student self-ratings determining measures of English listening and speaking competency of BCSE 2008 graduates against the prescribed indicators.

The study explored the relationship of CA marks with the BCSE 2008 exam marks relative to the relationship with the teachers' ratings and student self-ratings. Correlation research '…allows researchers to determine not only whether a relationship between variables exists, but also the degree of the relationship between them' (Gall, Gall and Borg, 1999: 211).

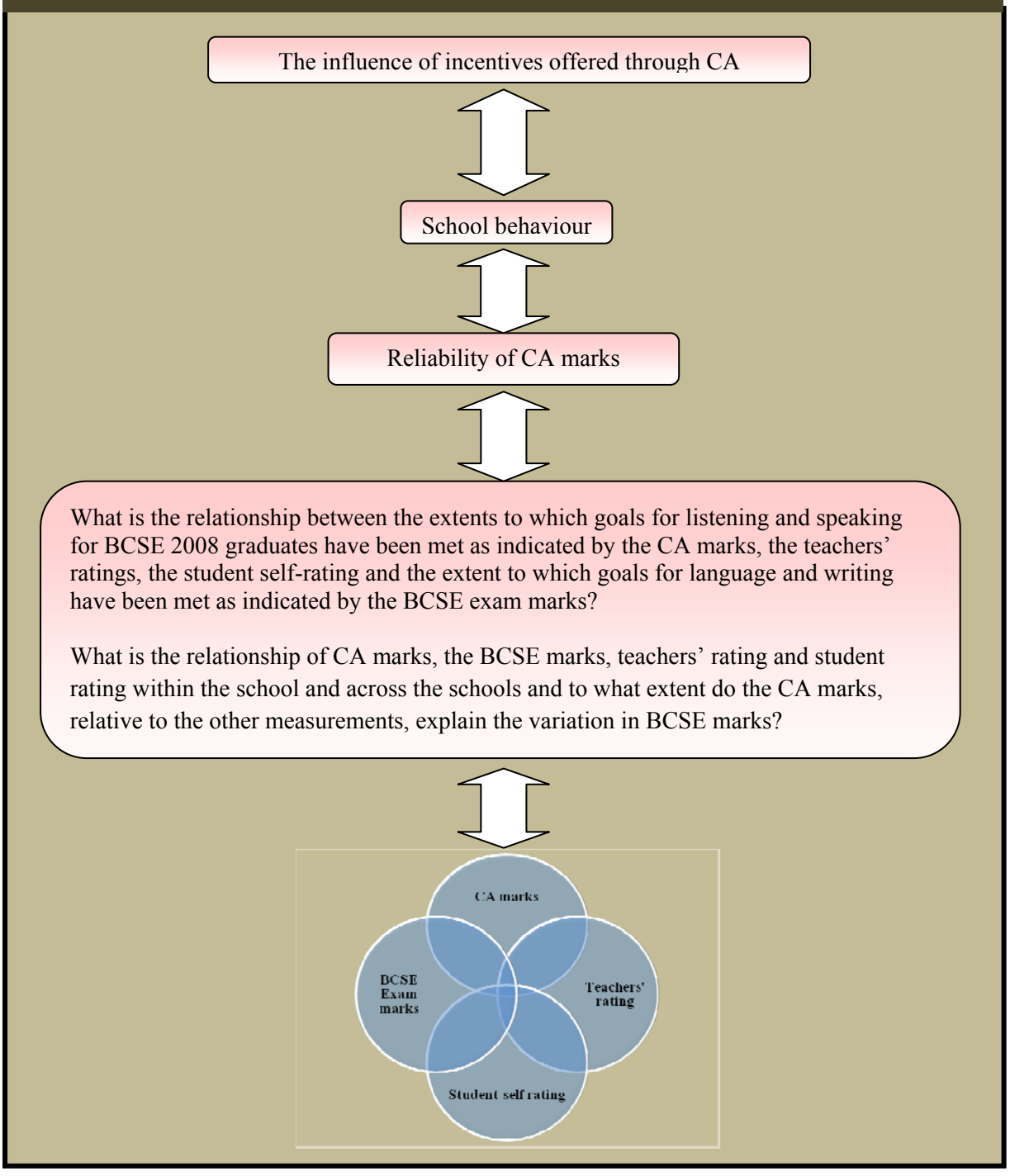
The relationships of the four variables are explored to address the research questions:

1. *What is the relationship between the extents to which goals for listening and speaking for BCSE 2008 graduates have been met as indicated by the CA marks, the teachers' ratings, the student self-rating and the extent to which goals for language and writing have been met as indicated by the BCSE exam marks?*

2. *What is the relationship of CA marks, the BCSE marks, teachers' rating and student rating within the school and across the schools and to what extent do the CA marks, relative to the other measurements, explain the variation in BCSE marks?*

The exploration of the relationship of the four variables in figure 3, illustrates how reliability of the CA marks is operationally equated against its correlation with the BCSE marks, the teachers' rating and the student-self rating. The study of relationships addresses the research questions informing the reliability of the CA marks. The reliability of the CA marks, indicative of school and teacher behaviour, determines the influence of incentives offered through the decentralized school-based continuous assessment (CA) marks.

Figure 2: Design



6.2.1 Internal validity

It is noted that ensuring internal validity, defined as credibility of the design where the observations can be attributed to the variables under study by controlling for the influence of all extraneous variables, is applicable to studies of relationship (Drew, Hardman and Hosp, 2008).

In exploring the reliability of the CA marks by studying its relationship with the BCSE marks, the study employs measurements of teachers' and student ratings for triangulation of relationships to validate the results of the study.

To control for the Hawthorne effect, the teachers and students were not informed of the correlation purpose their ratings would serve. This information was withheld intentionally to control for intentional conformity of teachers' and student ratings with the CA marks. The Hawthorne effect is defined as "a change in sensitivity or performance by the participants that may occur merely as a function of being in an investigation" (Drew et al. 2008, p. 223).

7 Methodology

In presenting the methodology, section 7.1 describes and explains the procedures followed in determining the sample and explaining the limitations associated with its generalizability. A description of the measurements and instrumentations used are presented in section 7.2 along with discussions on their validity and reliability. The data collection procedures and the methods of analysis are discussed in sections 7.3 and 7.4 respectively. The limitations of the study are explained in section 7.5 and section 7.5 is a note on the ethical considerations of the study.

7.1 Sample:

A total number of 7982 class 10 students (4143 boys and 3839 girls) appeared the Bhutan Certificate of Secondary Education (BCSE) 2008 examinations from 61 schools (36 middle secondary schools and 25 higher secondary schools including two private schools). Of the total students who appeared the examination, 7526 (94.29%) students successfully graduated secondary education through certification of the BCSE.

Of the 7526 students who graduated, the 3233 students who met the required overall cut-off score of 62.2% for admission to class 11 in the 25 government higher secondary schools comprises the sample population. The overall cut-off score of 62.2% for 2009 is an aggregate of 311 marks in English and four best subjects (Ministry of Education, Bhutan Board of Examinations notesheet, February 9, 2009). Due to foreseen constraints in locating BCSE 2008 graduates who are not in schools and owing to the time limit for data collection, this criterion has been imposed on the study. The criteria restricts the sample to only the accessible population that met the entry level cut-off criteria set by the Ministry of Education, Bhutan. It is noted as a limitation of the generalizability of the findings of the study.

Multistage sampling was followed in determining the sample. A list of the 25 higher secondary schools was obtained from the Bhutan Board of Examinations. Of the 25 higher secondary schools, a stratified selection of 10 schools was made. The sample units of 10 higher secondary schools represent 40% of the population of higher secondary schools. Trochim (2005) notes, that a stratified random selection allows for representation of the key sub groups of the population. The sample was selected based on location of the school. Region representation is taken into consideration. This is important as the four regions differ in terms of their ethnic composition, language spoken and their location, i.e. whether urban or semi-urban. To ensure fair representation of the regions, the number of schools selected for the regions was based on the total number of schools and students in the region. The sample has 2 higher secondary schools from the Central region, 3 from the Eastern region, 2 from the southern region and 3 from the Western region.

From each school selected as a unit, all the 2008 class 10 English teachers who were still in the same school and 15 of their students in 2008 who were admitted to class 11 comprise the sample cases. The 15 students for each teacher were selected through systematic random sampling, except in those cases where choice was not an option.

7.1.1 Limitations of sample

The sampling model, specific to the year 2008, is not representative of other cohorts of BCSE graduates. Thus the results of the study and any inferences drawn may be implied only with reference to the year 2008.

The sample selection for the survey allows for generalizability only to that population of BCSE 2008 graduates who met the cutoff point of 62.2%. The survey measured teachers' and students-self ratings and hence the imposition of the findings of the study with reference to the two measurements on the entire population of the 2008 BCSE graduates is not valid. However, data available in the form of the BCSE exam marks and CA marks for the population of students who appeared the 2008 BCSE examination may serve to allow for some generalization to the entire population.

7.2 Measurements and instrumentation

The measurements in the study are:

- teacher rating of students for English listening and speaking competencies, summarised as a total score for each student;
- student self-rating for English listening and speaking competencies, summarised as a total score for each student;
- the BCSE 2008 Examination raw marks for English paper I, represented by a single score for each student and;
- the CA marks submitted to the BBE by schools for student performance in English listening and speaking competencies, represented by a single score for each student.

7.2.1 Teacher questionnaire and student questionnaire

The survey focuses on determining a measure of the level of achievement of BCSE graduates' English listening and speaking skills against the indicators for levels of achievement as reflected in level 7 and 8 of the 'Silken Knot'. Two instruments were developed for the measurements; the teacher questionnaire and student questionnaire.

The questionnaires comprise teachers' assessment of students and student-self assessment of English listening and speaking competencies against the indicators of levels of achievement. A total of thirteen indicators for levels of achievement; seven from level 7 and eight from level 8 were incorporated in the development of the rating items in the questionnaires. The Silken Knot prescribes the indicators in level 7 and in level 8 as corresponding to expected performance of students in classes 9 through to12. A review of the learning outcomes for listening and speaking for class 10 as stated in the 'BCSE English Teacher's Guide' (CAPSD, 2005), confirmed the integration of both level 7 and 8 indicators.

The teacher questionnaires (Appendix A) comprise two sections; section A with eleven items on the general information on teacher characteristics, teaching practices and conditions and section B with thirteen rating items which is a direct translation of the thirteen indicators for level 7 and 8.

The student questionnaire (Appendix B) consists of four sections; A, B, C and D. The thirteen items in section A gather general information on student characteristics such as, gender, background and teaching practices and section B, C and D comprise rating items. Items 1 to 31 in section B are based on the national standards for listening and speaking competencies of BCSE graduates; items 32 to 40 in section C relate to writing, language and, reading competencies of the BCSE graduates; items 41 to 43 in section D comprises items on academic self-concept taken from the OECD's Brief Self-Report Measure of Educational Psychology's Most Useful Affective Constructs: Cross-Cultural, Psychometric Comparisons Across 25 Countries by Marsh, Hau, Artelt, Baumert and Peschar (2006).

7.2.1.1 Validity and reliability of teachers' and student-self ratings

To ensure the construct validity of the measurements obtained from the instruments, the following measures were followed in developing and finalizing the instruments:

- Both the teacher and student questionnaires were based on the same indicators of levels of achievements reflected in the standards for listening and speaking in the 'Silken Knot'.
- Both questionnaires were piloted in Bhutan. The teacher and student questionnaires, respondent feedback forms, guidelines for conduct of pilot and data compilation formats were developed and piloted in Bhutan with the assistance from the Bhutan Board of Examinations, Ministry of Education. Three higher secondary schools, not selected as units for the study, were identified for the pilot of the instruments. 3 teachers, one from each school and a total of 45 students, 15 taught by each teacher, were identified and administered the questionnaires.
- Both the questionnaires follow Likert response format using a four point Likert scale where:
  1 = Not true at all
  2 = Not true
  3 = True
  4 = Very true

The four point response scale ensures that there is not provision for a neutral response. Trochim (2005) notes, "… the respondent has to agree whether he or she is in agreement or disagreement with the item" (p. 112).

An example is provided in table 2 of the rating items of the student and teacher questionnaires. The example illustrated is of the first rating items from both questionnaires. The complete rating forms in the teacher questionnaire and student questionnaire can be referred to in appendix A1 and A2.

Table 2: Sample rating items from student and teachers questionnaires

Rating item from student questionnaire, section B

1. When a person speaks to me ***in English***, I am able to continue the conversation ***in English***.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐

Rating item from teacher questionnaire, section B

1. Match their talk to the demands of different circumstances.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐

Reliability of the instruments was ensured through measures of internal consistency and inter-item correlations of the pilot study and by taking the findings into account in finalizing the instruments.

- An item analysis using 'Quest' based on the Rasch Model showed an internal consistency of 0.96 for the Teacher Questionnaire and 0.89 for the Student Questionnaire (Appendix C).

- Internal consistency indicates how consistently the different items measure the same variable (Litwin. M. S., 1995). Cronbach's Alpha of 0.943 for the Teacher Questionnaire and 0.974 for the Students Questionnaire was reported(Appendix D). Cronbach's Alpha "… is a statistic that reflects the homogeneity of the scale" (Litwin. M. S., 1995, p. 24).

- Inter-item correlation analysis using SPSS indicated that item 3 in the Student Questionnaire did not have positive correlation with the other items. This was substantiated by item analysis using Quest which indicated also a misfit of item number 3 of the Student Questionnaire. This item was removed from the Student Questionnaire. Item 25 of the Student Questionnaire was also removed as a review of the items against the indicators showed that the item did not relate well to the indicator 7 which it was intended to measure. Items 4 and 42 of the Student Questionnaire also showed very poor and negative correlations with other items. A study of the items revealed that both the items appeared to stress on the situation rather that the specific skill being addressed. Both items were finalised after revisions.

7.2.2 Secondary measurements:

The BCSE 2008 examination results for English are reported as a single score for student performance in all English language skills assessed separately as English paper I and English paper II. Examinations for both papers are conducted and evaluated separately, English Paper I testing writing and language and English Paper II testing reading and literature. Performance in each paper is marked over 100 and converted to 80% and the CA marks from school accounting for the remaining 20% for each paper.

The BCSE exam marks retrieved from the BBE database refer to the raw marks attained by students for their performance in the examination for English paper I, prior to standardization of marks. The CA marks retrieved from the same database, correspond to the marks sent by the schools for student performance in English listening and speaking competency to inform the aggregate performance in English paper I.

Secondary data required for the study in the form of the CA marks for listening and speaking and the BCSE 2008 examination raw scores for English I were retrieved for the entire population of students who appeared the 2008 BCSE examinations.

7.2.2.1 Validity and reliability of the BCSE exam marks

Test development at the BBE incorporates procedures such as use of test specifications, item analysis, item banking, moderation and panelling of items which enhance the validity of the examination.

The English I question paper, testing student competency in writing and grammar, is developed based on the test specification which takes into account the learning outcomes to be measured, difficulty level of the task, the distribution of marks and time. The items of the question paper are also studied for test bias at the item level.

The question paper for English I comprises 13 multiple choice items (MCI), 22 short answer items (SAI) and 2 extended response items (ERI). The paper is organised into two sections, A and B. Section A, worth 40 marks comprises two ERIs which test writing skills. Item 1, worth 25 marks requires students to write an essay and item 2, worth 15 marks test students' letter writing skills. Section B with a total weighting of 60 marks comprises two subsections. The subsection on nature of language with 3 MCIs and 2 SAIs is based on student knowledge of the 'nature of language' and is worth 10 marks. The subsection on grammar has 10 MCIs and 20 SAIs testing students' application of grammar.

The number of items, their distribution across the difficulty levels, the marks they carry is presented in the sample blueprint in table 3, where:

- Q refers to question number,
- within ( ) are the item sub-numbering and
- within [ ] are the marks allocated,
- the last column indicates the total marks for each skill area assessed and the last row reflects the distribution of marks for each level of difficulty.

Table 3: Sample table of specifications

Difficulty / Skill	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
Section A Essay writing					Q1 [25] ERI		**25**
Section A Letter Writing			Q2 [15] ERI				**15**
Section B Nature of Language	Q1(i)-(ii) [1x2] MCI	Q1(iii) [1] MCI					**10**
	Q2 (i) (3) SAI			Q2(ii) [4] SAI			
Section B Grammar and Structures		1(i)-(vi) [1x6] MCIs	1(vii)–(x) [1x4] MCIs				**10**
				2(i)–(x) [1x10] SAIs			**10**
				3(i)-(v) [1X5] SAIs			**5**
						Q4 (1x5) SAI	**5**
Total	**5**	**7**	**19**	**19**	**25**	**5**	**80**

The reliability of the BCSE examination marks for English may be assumed in consideration of the controlled marking system followed during the evaluation of the student responses. Approximately 45% of the total population of the class 10 English teachers participate at the camp organised for the evaluation of student performance in English in the BCSE examination. The BBE criteria for the selection of markers are followed in the selection of teachers for the marking camp. A standardization procedure is undertaken prior to marking to discuss and adopt the marking criteria and the standards of the levels of performance and to ensure inter marker reliability. Despite the additional costs incurred, double marking system is utilised for the marking of the essays as it was found to increase the reliability of the marks awarded for the performance in extended responses.

The marks for English paper I, retrieved from the BBE database are the raw marks, i.e. the actual marks awarded received prior to standardization. Based on the system and procedures followed in the setting of questions and in the evaluation of the BCSE English I examination paper, the study assumes that the marks thus retrieved are both valid and reliable.

The Continuous Assessment (CA) marks for English paper I based on listening and speaking skills, submitted by the school to BBE, are scored against a total of 20 marks. The marks were retrieved from the BBE database. The CA marks are neither moderated nor standardized at the BBE and are merely added to the marks scored in the examination to arrive at an aggregate score for performance in English paper I. The CA marks collected are the exact marks submitted by the schools over a total score of 20. The CA marks for listening and speaking is the focus of the study.

7.3 Data collection procedures

The procedures followed are reported corresponding to tasks undertaken in preparation for, and during data collection.

After having finalised the questionnaires and made a stratified selection of the 10 higher secondary schools, the information on the study was conveyed and the approvals required to collect data was sought from: the Secretary of Examinations, The Bhutan Board of Examinations (BBE), for use of BCSE 2008 examination results; and the Director of School Education, Ministry of Education, for administration of teacher and student questionnaires in the 10 higher secondary schools selected for the study at the proposed time (Appendix E 1).

The following activities were undertaken during actual data collection in Bhutan:

- Principals of selected schools informed of the study over phone and date for admission of class 11 students confirmed.
- Formal letter of information on study with copy of letter from the Director of School Education faxed to all 10 higher secondary schools. Principals requested for the appointment of teacher to assist in data collection, hereafter referred to as research assistants (Appendix E 2).

- From all schools, according to the instructions (appendix A5), a list of the BCSE 2008 graduates of the school who had registered for admission to class 11 were compiled and forwarded by the research assistants.
- From this list, students were selected randomly taking into consideration the gender of the population of the school's BCSE graduates.
- The researcher administered the questionnaires in four schools within the four dzongkhags (districts) which were closer and the research assistants of the other six dzongkhags administered the questionnaires in the own schools. The duly completed questionnaires were forwarded by the research assistants.

To ensure consistency of the data collection procedures:

- a list of instructions to be followed consistently in all schools for the administration of the questionnaires was developed and sent to the six teachers appointed as research assistants (appendix A5).
- the instructions were followed consistently by both research assistants and the researcher at the time of data collection and communication over phone took place regularly with the research assistants to ensure this consistency.

7.4 Data analysis

The analysis was done at the student level with marks for each student against each measurement constituting the unit of analysis.

The BCSE exam marks and the CA marks are represented by independent marks for each student. The analysis was based on the summary measures of teachers' assessment and student-self assessment. The ratings were summarised based on the following computations:

The teachers' rating on the 13 rating items was summarized as a total score for each student. The conversion of responses to scores was based on the predetermined and pre-tested scale of 1-4 where: the response not true at all = 1; not true = 2, true = 3 and very true = 4. The teachers' rating was thus computed over a total weighting of 52. Averages were computed for rating items on the student questionnaire corresponding to the indicators reflected in teachers' questionnaire. The 13 averages of the student self-rating thus computed were summarised as a total score for each student over a total score of 52.

The summary measurements of teachers' ratings and student self-ratings were used for the analysis along with the BCSE exam marks and the CA marks.

7.4.1 Overall analysis of relationships

To study the relationship between '*the extent to which goals for listening and speaking for BCSE 2008 graduates have been met as indicated by the CA marks, the teachers' ratings, the student self-rating*

and the extent to which goals for language and writing have been met as indicated by the BCSE exam marks?' the following analyses were performed using SPSS:

Descriptive analyses on the four variables to study the central tendencies and dispersion of the measurements. To equate the measurements on a common scale for better comparison of the descriptive information provided, the measurements for the descriptive analysis are based on the marks converted to the percent scale.

Correlation analysis was performed to study the direction and degree of relationship among the four variables. To address the hypotheses and research question, the correlation analysis was performed as a single undifferentiated group level to study the overall relationship of the CA marks with the other measurements.

The correlation analysis was also undertaken to explore the relationship of other language skills, viz. student writing, grammar, reading competencies and academic-self concept with the CA marks, teachers' rating, student self rating and the BCSE exam marks. These measures correspond to student-self rating in the last 12 items of the questionnaire described in 7.2.1 under subtitle 'Teacher questionnaire and student questionnaire'. For the analysis of the last 12 items of the student rating form, averages were estimated for three tasks corresponding to student-self assessment of writing, grammar, reading and literature and their academic-self concept respectively. The estimated marks were used to study relationships of the student ratings on the four areas with the CA marks, teachers' rating, student-self rating and the BCSE exam marks.

7.4.2 Analysis of relationship within schools and between schools

To explore the second question focusing on variation within and between schools correlation analysis was performed at the school level. School means computed for each measurement, were correlated to study for variations between schools. Student marks in each measurement relative to the school means were computed to allow the study of relationships of the measurements within the schools. The computations of the marks were made using the aggregate and compute-variable functions in SPSS.

7.4.3 Regression analysis to study relationship of variation in measurements

Further exploration of the relationship of the four measurements was undertaken using multiple regressions to investigate how well the CA marks explain the variation in BCSE exam marks relative to teachers' rating and student-self rating as predictors. The BCSE mark was taken as the outcome variable to be predicted by the CA marks, teachers' rating and student-self rating taken as independent variables.

The model used in the analysis may is illustrated below where, CA marks, teachers' rating and student rating are included as predictors in hierarchical order:

Variation in BCSE marks = $\beta_0 + \beta_1$ CA marks$_1$ + β_2 Teacher's rating$_1$ + β_3 Student rating$_1$ + ε_i

7.5 Limitations of the study

In addition to teachers' rating and student-self rating, the BCSE exam mark is one of the standard reference points based on which inferences on the accuracy of the CA mark is drawn. Though procedures undertaken by the BBE to ensure the validity and reliability of the examinations are described, issues pertaining to comparability require in depth investigation beyond the scope of this study. Also since the survey is a onetime measure of students' English listening and speaking competencies, the measurement is at best valid only with reference to the specific BCSE 2008 cohort population represented by the sample. Thus caution in inference of the findings beyond the scope of the study to indicate trends over time is advised, in consideration of comparability issues of the examinations over years and for different cohorts.

The study will be judging the perception of class 10 graduates and their teachers after a break of two months. This implies the study may have limitations in that both the teacher respondents and students are required to recall experiences to be able to accurately rate the students and themselves respectively. However, that no teaching/learning has taken place during the break period reduces the effects of break in time.

Though results were obtained for the population of students who appeared the BCSE examinations based on the BCSE exam marks and the CA marks, the survey sample to determined the teachers' rating and student-self rating limit generalizations to the sample population. Based on an exploration of relationships relative to the teachers' rating and the student-self rating, the study does not allow for direct inference from the results of the data available for the entire population to be conclusive. The findings of the study are generalizable to the sample population of the BCSE graduate who met the prescribed cut-off over and above successful completion certified by the BCSE and secured admission into class 11 in higher secondary schools.

7.6 Ethical considerations

"An omission deception could mean that an investigator does not fully inform participants about important aspects of the study" (Drew, Hardman and Hosp, 2008). In conducting the survey, to derive teachers' rating and student self-rating for listening and speaking competencies against which relationship studies for the CA marks could be undertaken, respondents were not informed of the relational purpose the ratings would serve. This information was withheld to avoid confirmatory behaviour of the respondents in their ratings through the influence of the knowledge of the CA marks. The concern for truthful answers and validity of data may necessitate the use of deception with due consideration of ethical implications (Trochim, 2005).

8. Results

The results of the analysis are presented in two sections corresponding to the analysis undertaken to address the two research questions.

Section 8.1 reports the results of the analysis to study the reliability of the CA scores relative to its relationship with the other measurements. Section 8.2 presents results of the second question, viz. the differences within schools and between schools and the results of the regression analysis exploring the extent to which the CA marks, the teachers' rating and the student-self rating explain variation in the BCSE exam marks.

The results of the analysis for the CA marks and BCSE exam marks, available for the population of the students registered for the BCSE 2008 examination follow the results of the study in both sections.

8.1 Results of the Analysis of relationships

Results pertaining to the study of the relationship of the measurements are reported in four subsections. 8.1.1 presents the central tendencies and dispersions of the four measurements; 8.1.2 reports results of the overall correlations of the four measurements; 8.1.3 presents a summary of the correlation results to test the hypotheses and 8.1.4 reports correlations of student-self rating for competency in writing, grammar, reading and academic self concept in relation to the four measurements.

8.1.1 The central tendencies and dispersions of the four measurements

As explained under 7.4.1, the BCSE marks and the CA marks scored over a total of 80 and 20 marks respectively were equated on the percent scale. The calculations were computed using the aggregate function in SPSS.

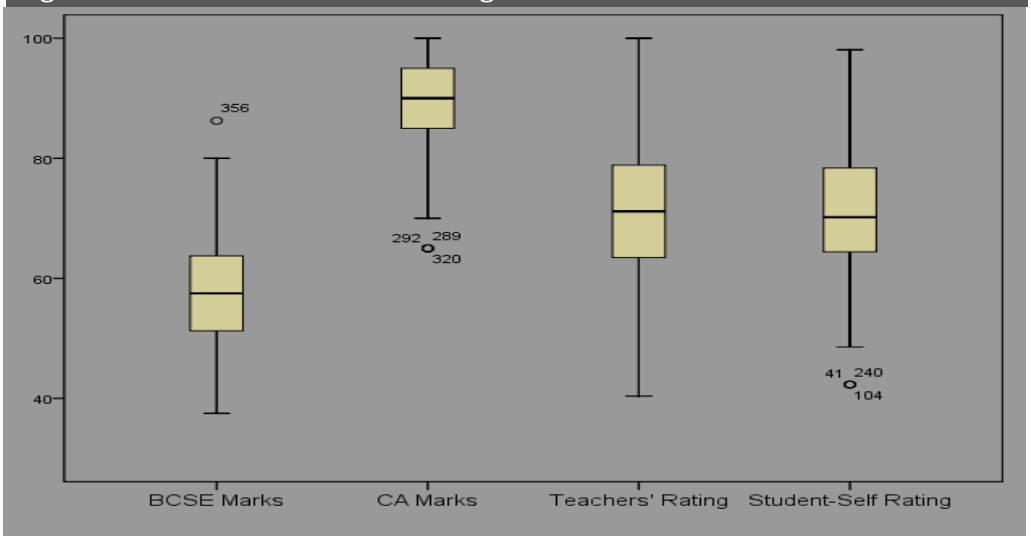
Table 4: Central tendencies and dispersions

		BCSE marks	CA marks	Teacher Rating	Student-Self Rating
N	Valid	365	365	365	365
	Mean	57.7089	87.7397	70.7795	71.0840
	Median	57.5000	90.0000	71.1500	70.1900
	Std. Deviation	8.62270	7.86393	11.91846	10.15744
	Skewness	.275	-.470	.072	.205

The central tendencies and dispersion of the variables as shown in Table 4 indicate significant difference in the mean marks of the variables. CA marks have the highest mean (87.74 with a standard deviation of 7.86) compared to BCSE 2008 examination marks (57.71 with a standard deviation of 8.62), teacher rating (70.78 with a standard deviation of 11.91), student self-rating (71.08 with a standard deviation of 10.16) and. The comparatively higher values of the median and the mean with corresponding lower values for the standard deviation of the CA marks for listening and speaking confirm the distribution of the marks a being concentrated towards the higher scores. This distribution is confirmed by the corresponding negative value for skewness (- .470) of CA marks. The direction of skewness for all other variables is positive.

The box plots in figure 4 illustrate the results of the distribution of the marks and indicate the median of the four measurements. The interquartile range and the median for the CA marks for listening and speaking is located comparatively much higher on the y axis reflecting the skewed distribution of the CA marks. The top quartile for the CA marks is much smaller than its bottom quartile and interquartile range indicating less discrimination in the award of higher scores.

Figure 4: Distribution of CA marks against other variables



The results of the analysis of the data for the population of the BCSE 2008 candidates are presented in table 5. Results of the population are similar to that of the survey with consistently higher values for the CA marks for listening and speaking with respect to the mean and median and lower standard deviation compared to the BCSE examination marks. The distribution of the CA marks for listening and speaking of the population is also negatively skewed, indicating concentration of marks towards the higher scores.

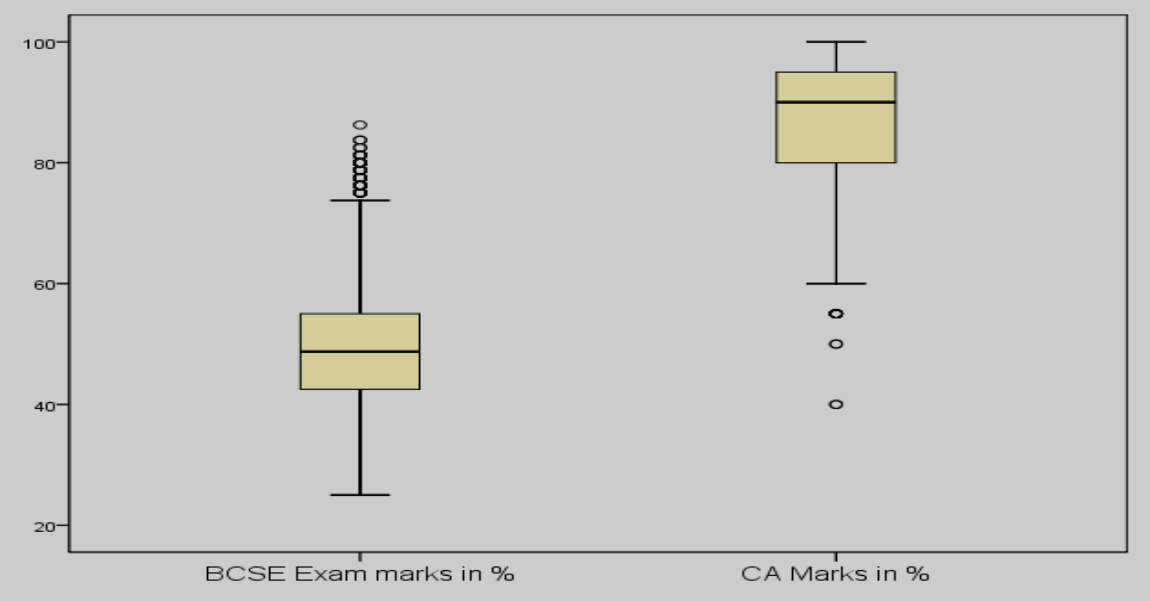
Table 5: Central tendencies and dispersions for Population

		BCSE marks	CA marks
N	Valid	7982	7982

	Mean	49.348	87.301
	Median	48.750	90.000
	Std. Deviation	9.2095	8.2239
	Skewness	.636	-.653

Figure 5, illustrating the box plots of the BCSE marks and the CA marks for the population of the BCSE candidates, shows concentration of the CA marks towards the higher scores compared to the BCSE marks. The location of its median and its interquartile range are very far from that of the BCSE 2008 examination marks on the y axis. The comparatively shorter length of its top quartile reflects less discrimination in the award of higher scores.

Figure 5: Distribution of CA marks and BCSE marks for Population



The results of the descriptive analysis of the CA marks and the BCSE marks for the population of BCSE candidates confirm the results of the study.

8.1.2 Results of overall correlation of measurements

The results are based on the correlation analysis of the individual students' marks on each of the four measurements. The correlation analysis was performed using the raw marks scored against each measurement; reported over 80 for BCSE, over 20 for CA and over 54 for both the teachers' and student-self ratings. Pearson's correlation coefficient 'r' is used to report the results of the correlation.

Table 6 reports the outputs of the correlation analysis. Though positive, the size of the correlations across all variables is low.

The CA marks share highest correlation with the teachers' rating with a coefficient of r = + .264, significant at p (1 tailed) < .001. Compared to its correlations with the BCSE exam marks with r = + .132,

p (1 tailed) < .05, the CA marks show hardly any correlation with the student rating in terms of the degree of the correlation. The degree of correlation between the CA marks and the student rating is statistically not significant.

Table 6: Correlation of CA marks with three measurements

		BCSE Marks	CA Marks	Teachers' rating	Student rating
CA Marks	Pearson Correlation	.132**	1.000		
	Sig. (1-tailed)	.006			
	N	365	365.000		
Teachers' rating	Pearson Correlation	.313**	.264**	1.000	
	Sig. (1-tailed)	.000	.000		
	N	365	365	365.000	
Student rating	Pearson Correlation	.297**	.034	.331**	1.000
	Sig. (1-tailed)	.000	.257	.000	
	N	365	365	365	365.000

**. Correlation is significant at the 0.01 level (1-tailed).

The results of the correlation analysis between the CA marks and the BCSE exam marks for the survey are corroborated by the results of the population for the same measurements. The correlation of the CA marks and BCSE exam marks for the population of BCSE candidates, though slightly higher than that of the survey is still low, with r = + .238, p (1 tailed) < .001.as reported below in table 7.

Table 7: Correlation of CA marks with BCSE marks for Population

		BCSE exam marks	CA marks
CA marks	Pearson Correlation	.238**	1.000
	Sig. (1-tailed)	.000	
	N	7982	7982.000

**. Correlation is significant at the 0.01 level (1-tailed).

8.1.3 Summary of results testing the hypotheses

This sub-section summarizes the results of the overall correlation to address the hypotheses stated under 4.5.1 in exploration of the first question.

Results testing hypothesis 1: *CA marks will have stronger positive correlation ≈ +.7 with teacher rating of student performance.* The CA marks for listening and speaking is positively related to teachers' rating with a coefficient of r = .264, significant at p < .001. Thus, though teachers' rating reports the strongest positive correlation with CA marks than the other measurements, the result do not support the degree of the relationship hypothesized.

Results testing Hypothesis 2: *CA marks will have strong positive correlation ≈ + .3 with student rating.* The correlation between student rating and the CA marks with r = + .034, is not significant at p > .5. The low significance of the correlation indicates there is over 25 percent probability that the negligible positive relationship is a chance occurrence. This may be interpreted as implying no relationship between the CA marks and the student-self rating.

Results testing Hypothesis 3: *CA marks will have strong positive correlation ≈ + .6 with BCSE exam marks.* The CA marks for listening and speaking though positively related to BCSE 2008 Marks has a very low coefficient of r = + .132, significant at p < .05. The hypothesis cannot be accepted with respect to the degree of the relationship predicted.

Results testing Hypothesis 4: *There will be statistically significant positive relationships among the four measurements.* Of the results, the highest degree of positive correlation is reported between the teacher rating of students and the student-self rating with a coefficient of r = + .331, significant at p < .001. Teachers rating also shows positive correlation with the BCSE 2008 Marks with a coefficient of r = + .313, significant at p < .001. The student-self rating though positive has a lower correlation with the BCSE 2008 marks with a coefficient of r = + .297, significant at p < .001.

Two important observations to be noted from the test of the fourth hypothesis are;

- Teachers' rating reportedly sharing the strongest correlation with all other measurements.
- Better agreement of teachers' rating and student-self rating with the BCSE marks than with the CA marks.

The results appear to confirm the hypotheses on the direction of the relationship but do not allow the acceptance of the strength of the relationship hypothesized. However, there appears to be some agreement with the hypotheses with regard to the comparative strengths of the correlations hypothesized. In agreement with the hypotheses, the correlation of the CA marks is strongest with that of the teachers' rating, followed by its correlation with the BCSE marks.

8.1.4 The relationship of other language skills and student academic-self concept with the measurements.

The marks estimated for student-self ratings for writing, grammar, reading and literature and their academic-self concept were correlated to explore the relationship within the measurements itself as well as their relationship with the CA marks, teachers' rating, student-self rating and the BCSE marks.

The results of the correlation analysis are presented in two separate tables. Table 8 presents the results of the correlation of the student-self ratings on writing, grammar, reading and academic self concepts and table 9 reports the correlations of the ratings on the three language skills and academic self concept with the four measurements of the study.

Table 8: Correlations between student ratings of three language skills and academic-self concept

		Student Rating for Writing	Student Rating for Grammar	Student Rating for Reading and Literature	Student Rating for Academic Self-Concept
Student Rating for Grammar	Pearson Correlation	.560**	1.000		
	Sig. (2-tailed)	.000			
	N	365	365.000		
Student Rating for Reading and Literature.	Pearson Correlation	.481**	.587**	1.000	
	Sig. (2-tailed)	.000	.000		
	N	365	365	365.000	
Student Rating for Academic Self-Concept	Pearson Correlation	.468**	.391**	.400**	1.000
	Sig. (2-tailed)	.000	.000	.000	
	N	365	365	365	365.000

The correlations between the different language skills and academic self concept reported in table 8, are positive and significant with the size of the correlation ranging between + .40 and + .59. Student rating on academic-self concept appears to share the strongest relationship with student rating for writing with $r = + .47$, $p < .001$. Overall, moderate correlations are indicated between the student ratings for different skills in English and academic-self concept.

Table 9: Correlation of student ratings in different language skills with four measurements

		Student Rating for Writing	Student Rating for Grammar	Student Rating for Reading & Literature	Student Rating for Academic Self-Concept
BCSE Marks	Pearson Correlation	.310**	.325**	.175**	.239**
	Sig. (2-tailed)	.000	.000	.001	.000
	N	365	365	365	365
CA marks	Pearson Correlation	.076	.027	.048	.058
	Sig. (2-tailed)	.146	.602	.362	.269
	N	365	365	365	365
Teacher Rating	Pearson Correlation	.206**	.122*	.162**	.220**
	Sig. (2-tailed)	.000	.020	.002	.000
	N	365	365	365	365
Student Rating for Listening & Speaking	Pearson Correlation	.613**	.568**	.563**	.465**
	Sig. (2-tailed)	.000	.000	.000	.000
	N	365	365	365	365

In table 9, moderate to strong relationships are reported between student ratings for listening and speaking and their ratings for writing, grammar and reading skills with correlations ranging between + .61

and + 0.47, significant at p < .001. The results indicate that student-self rating for listening and speaking shares the strongest relationship with writing and grammar competencies (r = + .61 and + .57 respectively) which are the corresponding skills assessed in English paper I though the BCSE examination. The conformity of the student-self rating for listening and speaking to their ratings for competency in the three language skills and academic concept also imply consistency of student-self ratings to a certain degree.

The result in table 9 also indicates that student self-rating for the three language skills and academic-self concept have stronger positive relationship with the BCSE exam marks than with the CA marks and teachers' ratings. Correlation coefficients with the BCSE exam marks range between + .33 and + .24, while the correlations with CA marks are not significant and range between + .08 and + .03. This may be taken to imply that a student's perception of his/her language skills and academic-self concept appears to be better informed by performance in the BCSE exam than by the CA marks given by the teacher.

8.2 Results of differences within schools and between schools and the regression analysis

Sub section 8.2 presents results of the second question. Results of the differences in measurements within and between schools are reported in 8.2.1 and the result of the regression analysis is presented in 8.2.2.

8.2.1 Results of correlation within schools and across schools

The results of the correlation analysis based on the marks of the four measurements relative to the means of the school, to study relationship of measurements between schools are reported in table 10. The results of the same analysis for the population of students who appeared the BCSE 2008 examination are presented in table 11 for the CA marks and BCSE marks.

Table 10: Correlations of measurements within schools (relative to means of school)

		BCSE exam marks	CA marks	Teacher ratings	Student rating
CA marks	Pearson Correlation	.319**	1.000		
	Sig. (1-tailed)	.000			
	N	365	365.000		
Teachers' ratings	Pearson Correlation	.372**	.222**	1.000	
	Sig. (1-tailed)	.000	.000		
	N	365	365	365.000	
Student rating	Pearson Correlation	.322**	.090*	.364**	1.000
	Sig. (1-tailed)	.000	.043	.000	
	N	365	365	365	365.000

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

The correlation coefficients of the table 10 show that within schools the students with high BCSE scores tend to get relatively high CA scores, high teacher ratings and give high self ratings. The correlation between CA marks and BCSE marks within schools is moderate with r = + .319 significant at p < .001.

The results for the same analysis on CA marks and BCSE exam marks for the population of students who appeared the BCSE 2008 examination, presented in tables 11 report similar results. The correlation of CA marks and BCSE marks within schools is positively significant.

Table 11: Correlations of measurements within schools for Population

		BCSE exam marks	CA marks
CA marks	Pearson Correlation	.453**	1.000
	Sig. (1-tailed)	.000	
	N	7982	7982.000

**. Correlation is significant at the 0.01 level (1-tailed)

To study the relationship of the four measurements across the schools, the means of the four measurements for the 10 schools were correlated. In reporting the correlation coefficients between school means of the four measurements in table 12, the p values were calculated taking into account the 40% sampling fraction.

Table 12: Correlations of measurements between schools (between school means)

		School means of exam marks	School means of CA marks	School means of Teachers' ratings	School means of Student rating
School means of CA marks	Pearson Correlation	-.658*	1.000		
	Sig. (2-tailed)	.002			
	N	10	10.000		
School means of Teachers' ratings	Pearson Correlation	-.294	.613	1.000	
	Sig. (2-tailed)	.092	.003		
	N	10	10	10.000	
School means of Student rating	Pearson Correlation	-.174	-.253	.117	1.000
	Sig. (2-tailed)	.215	.126	.297	
	N	10	10	10	10.000

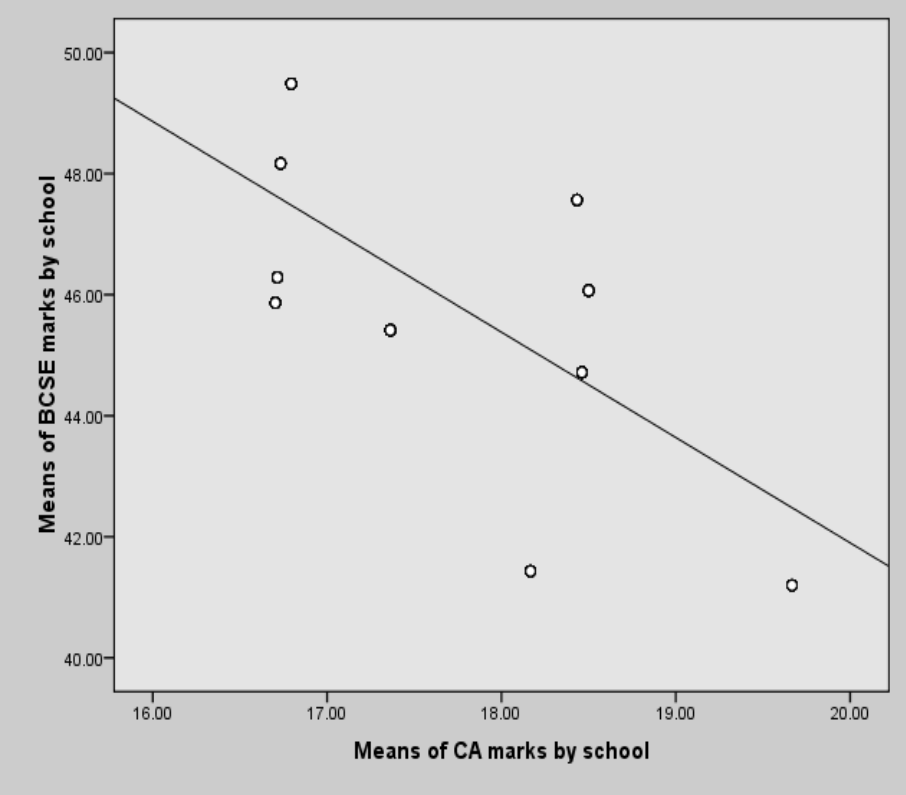
**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Of particular interest and concern is the significant negative relationship between the BCSE mean exam score of schools and CA mean score of schools, r = - .66, p < .05, meaning schools where the students on average score high on BCSE exam marks tend to score relatively low CA averages and vice versa.

The results of negative correlation of the CA marks and the BCSE exam marks are better elucidated through the scatter-plot in figure 6.

Figure 6: Scatter-plot of school means of CA marks and BCSE marks



8.2.2 Results studying relationship of variation in measurements

This subsection reports results pertaining to the regression analysis undertaken to explore the extent to which variations in the BCSE exam marks is explained by the CA marks, the teachers' rating and the student-self rating. The model used may be summarised as the following equation;

Variation in BCSE marks = $\beta_0 + \beta_1$ CA marks$_i$ + β_2 Teachers' rating$_i$ + β_3 Student rating$_i$ + ε_i

Table 13: Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.132[a]	.018	.015	8.55864	.018	6.469	1	363	.011	
2	.317[b]	.101	.096	8.19929	.083	33.515	1	362	.000	

3	.380[c]	.144	.137	8.01032	.043	18.281	1	361	.000	1.638

The model summary in table 13 explains the CA mark, teachers' rating and student-self rating as having been entered into the model as predictors hierarchically, where CA mark is entered first, followed by teachers' rating and student-self rating.

The R^2 statistic is a measure of the amount of variability in the outcome (BCSE marks) that can be explained by the predictors (CA marks, teachers' rating and student-self rating). The R^2 statistic indicates that the CA mark as predictor explains only 1.8% of the variation in BCSE marks. However, with the inclusion of teachers' rating as a predictor, 10.1% of the variation in BCSE marks can be explained by both the CA marks and teachers' rating. This implies that on its own teachers' rating can explain 8.3% of the variation in the BCSE marks.

Similarly, since the inclusion of student self rating as predictor in the model explains 14.4% of the variation of the BCSE marks, by itself the student rating explains 4.3% of the variation in BCSE marks. Thus, it appears that comparatively, the teachers' rating explains most of the variation in the BCSE marks, increasing the R^2 statistic of the model by .083. This is indicated by the R^2 change statistics. This change in the amount of variation explained by including teachers' rating is reported by the F-ratio of 33.515, significant at $p < .001$.

However, the model can explain only 14.4% of the total variation in the BCSE marks. This implies that 85.6% of the variation in BCSE marks not accounted for by the CA marks, the teachers' rating and the student rating may be influenced by other factors.

Table 14 shows improvement in the model in prediction of BCSE marks associated with inclusion of teachers' rating as predictor in the second model indicated by the F value of 20.282 significant at $p < .001$.

Table 14: ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	473.889	1	473.889	6.469	.011[a]
	Residual	26589.869	363	73.250		
	Total	27063.759	364			
2	Regression	2727.064	2	1363.532	20.282	.000[b]
	Residual	24336.695	362	67.228		
	Total	27063.759	364			
3	Regression	3900.092	3	1300.031	20.261	.000[c]
	Residual	23163.667	361	64.165		
	Total	27063.759	364			

The results of the coefficients of the model reported in table 15 indicate that both teachers' rating and student rating contributes to the model in predicting BCSE marks with $t = 4.155$ and 4.276 respectively, and significant at $p < .001$.

Table 15: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	44.978	5.025		8.951	.000	35.097	54.860
	CA Marks	.145	.057	.132	2.544	.011	.033	.257
2	(Constant)	37.267	4.995		7.461	.000	27.444	47.090
	CA Marks	.058	.057	.053	1.030	.304	-.053	.170
	Teacher Rating	.216	.037	.299	5.789	.000	.143	.290
3	(Constant)	26.632	5.477		4.862	.000	15.861	37.403
	CA Marks	.072	.055	.066	1.303	.193	-.037	.181
	Teacher Rating	.161	.039	.223	4.155	.000	.085	.237
	Student Rating	.188	.044	.221	4.276	.000	.101	.274

The values and the corresponding significance of the coefficients in the model, confirm that compared to the CA marks, the teachers' rating and student-self rating significantly predict BCSE marks and thus comparatively, better able to explain the variation in BCSE marks.

The standardised coefficients show that when teachers' rating increase by 1 standard deviation (11.91846), BCSE marks increased by .223 standard deviations. The standard deviation for BCSE marks is 8.62270, constituting an increase in BCSE marks by 1.9228621 marks (.223 x 8.62270). That is, every increase in teachers' rating by 11.91846 marks explains or accounts for an increase of 1.9228621 marks in the BCSE marks. When student rating increase by 1 standard deviation (10.15744), BCSE marks increased by .221 standard deviations. That is, an increase in student rating increases by 10.15744, explains an increase of 1.9056167 marks in BCSE marks. The extent to which variation in BCSE marks is explained by the predictors in the model is not very high and thus account for less than 15% of the total variation.

9 Discussion:

Based on the theory of the influence of incentives, offered through decentralisation of the setting of standards on school and teacher behaviour, this study explored the reliability of the continuous assessment (CA) marks relative to its relationship with the teachers' rating, student-self rating and the BCSE exam marks. It studied also the relationship of the measurements within and between schools and investigated for the influence of other factors by exploring the extent to which variation in BCSE marks could be explained by the other measurements. The discussion focuses on the evaluation and interpretation of the results.

Discussion will be based on the two main questions addressed in the study and presented in two sub-sections, 9.1 and 9.2 each of which comprises discussions on the evaluation and interpretation of the results.

9.1 Discussion on first research question

Hypotheses were formulate to answer and guide the first question, *What is the relationship between the extent to which goals for listening and speaking for BCSE 2008 graduates have been met as indicated by the CA marks, the teachers' ratings, the student self-rating and the extent to which goals for language and writing have been met as indicated by the BCSE exam marks?*

The results of the correlation analysis to test the hypotheses though indicating positive relationship, failed to confirm the strength of the relationship of the CA marks with the BCSE exam marks and with the teachers' rating. The relationship with the student rating was found not to be significant.

A correlation of + .132 between CA marks and BCSE marks is reportedly low; not only does it fall short of the relationship recommended at + .7 for application of moderation processes as indicated by earlier studies (Deale as cited in Tam, 1977), it also drops below the minimum benchmark of + .3 below which the study (Elley and Livingstoneas cited in Tam, 1977) illustrates ineffectiveness of moderation processes. The low correspondence of the CA mark and the BCSE exam marks is also confirmed by the results for the population of the BCSE 2008 candidates reporting a slightly higher positive correlation of +.238 which still falls short of the recommended size of relationship.

That despite being measures of the same construct of English language skills, the lack of or poor degree of agreement between the CA marks and the BCSE exam marks may at first be interpreted to imply that students who speak English well do not necessarily write well in English and vice versa. However, a more logical interpretation supported by literature discussed under 5.1 in the review, would be the difference in the form and purposes of teachers' assessments and examinations, where teachers' award of CA marks informed by the formative purpose may result in CA marks that differ in composition from the BCSE exam marks (Harlen, 2008). However, the validity of this inference is lost when considering the poor relationship between the CA marks and teachers' rating itself.

The rejection of the hypothesis on the strength of the relationship between the CA marks and the teachers' rating may be slightly difficult to explain, both being measures of the same students' listening and speaking skills in English by the same teacher. The findings may be taken to imply difference in teachers' summative assessment as being dependent on the purpose of the assessment. The implication of the design of the study and the influence of the break-in-time, where teachers' and student-self ratings were measured after a break of two months, does not hold when considering the higher correlation of the BCSE marks with the teachers' rating and student-self rating than with the CA marks.

However, the interesting observation is the stronger relationship of all the individual measurements with the teachers' rating than with the CA marks. The test of the fourth hypothesis report the highest degree of positive correlation between the teacher rating of students and the student-self rating with a coefficient of $r = .331$, significant at $p < .001$ and between teachers rating with the BCSE 2008 Marks with a coefficient of $r = .313$, significant at $p < .001$. This indicates that relative to the relationship with other measurements, the teachers' rating appears to be a more agreeable estimate of students' competency in listening and speaking skills. The central tendencies and dispersions of the measurements also indicate better conformity of the teachers' rating with the BCSE marks and the students' rating.

Findings from the correlation of the measurements with student-self assessments in writing, grammar, reading and academic-self concept appear to support the earlier interpretation of the disagreeability of the CA marks relative to the other measurements. The results indicate stronger positive relationship of the three language skills and academic-self concept with the BCSE exam marks than with the CA marks. The relationship of language skills and academic-self concept with teachers' rating reportedly better than that with the CA marks. Student self-assessments in writing, language, reading and academic self concept may be considered consistent and reliable estimates in that they correlate significantly with each other as well as with the student rating for listening and speaking, and to a degree higher than that observed among the four measurements of the study.

The results of the overall correlations indicate the anomalous relationship of the CA marks with the other measurements, relative to the relationships between the different measurements.

To address the first question, the results of the descriptive analysis and the relationships of the four measurements may be thus summarized. That the goals for listening and speaking appear to have been met more extensively by the BCSE 2008 graduates as indicated by the school-based continuous assessment (CA) marks awarded by the teacher than indicated by the teachers' rating of student achievement and the student-self ratings. That the goals for writing, grammar and reading as indicated by the BCSE exam marks appear not to have been met as extensively as the goals for listening and speaking as measured by the CA marks.

9.2 Discussions on second research question

Discussions in this section pertain to the results of the correlation analysis to explore relationships of the measurements within schools and between schools and the regression analysis to address the following question:

What is the relationship of CA marks, the BCSE marks, teachers' rating and student rating within the school and across the schools and to what extent do the CA marks, relative to the other measurements, explain the variation in BCSE marks?

Discussions on the evaluation and interpretation of results addressing the second question are presented first for results of the correlation analysis investigating differences within and between schools followed by the results of the regression analysis investigating the extent to which the CA marks, the teachers' rating and the student-self rating explain variation in the BCSE exam marks.

Results of the correlation of the CA marks, the BCSE exam marks, the teachers' rating and the student-self rating relative to school means indicate that within the same school, students who score high on the BCSE marks also score relatively high on CA marks, the teacher rating and the student-self rating. These results also appear to suggest that within the schools, there appears to be a moderate degree of conformity within the three measurements of listening and speaking skills (CA marks, teacher rating and student-self rating) and the writing, grammar and reading skills as measured by the BCSE exam marks. Overall, within schools the measurements appear to conform to each other. Results for the population of BCSE candidates, i.e., 7982 students from the 61 schools, also confirm that within schools, students who score high on BCSE exam marks get relatively high CA marks.

However within the schools, the better conformity of the teachers' rating and student-self rating to the BCSE marks as compared to their relationship with the CA marks, intended to measure the same skills, confirms the findings on the inconsistency of teachers' assessments in the award of CA marks. It must be noted that the degree of conformity relative to other measurements within schools is also higher for the teachers' ratings than the CA marks. Thus, even within schools, though the measurements appear to conform to each other, the degree of conformity is lowest for the CA marks.

Results of the correlations of school means of the four measurements (CA marks, BCSE exam marks, the teachers' rating and the student-self rating) indicate negative correlations of the CA marks with the BCSE marks with $r = -.624$. This implies that schools were students on average score high marks in the BCSE exam tend to score relatively low CA averages and vice versa. Similar findings are reported in studies (Himmler and Schwager, 2007) (Reeves, Boyle and Christie, 2001) (Willingham, Pollack and Lewis, 2002) (Wikstrom and Wikstrom, 2005) (Thomas, Madaus, Raczek and Smees, 1998) discussed under 5.1.5.1 of the literature review.

The results may be summarised as indicating moderate degree of conformity between the CA marks and the BCSE marks within individual schools and a higher degree of nonconformity between schools. That within schools, the CA marks are awarded consistently and conform, to a certain degree, to performance in the BCSE exams, may be interpreted as teacher assessing students consistently against some established criteria or standards.

The conformity of CA marks and BCSE exam marks within schools and the contradiction of the two measurements between schools may be interpreted as indicative of low inter-rater reliability or the lack of comparability of teacher judgements between different schools. In discussing comparability of judgements between teachers, Buchan (1993) discusses standardization and moderation procedures to ensure comparability of teachers' assessment within and between schools. In studying peer and teachers'

assessment of oral skills, Magin and Helmore (2001) note the inadequacy in reliability of single teacher ratings in assessing oral presentation skills. The inter-rater differences may further be inferred to imply differences between schools in either the interpretation or the application of the standards of the levels of achievement for listening and speaking. The BCSE exam being a standardised test, implication of different standards across schools does not arise. Similar findings of differences between schools in teacher assessments and standardised tests are reported in studies (Himmler and Schwager, 2007), (Reeves, et al, 2001) and (Wikstrom and Wikstrom, 2005) discussed in the literature review under 5.1.5.1.

The results of the regression analysis confirm the findings of the correlation analysis, with teachers' rating determined as a better predictor and explaining the variation in the BCSE marks to a greater extent compared to the CA marks. That the CA marks do not account significantly to explain for the variation of BCSE marks confirms findings related to the nonconformity or lack of conformity of the CA marks to the other measurements.

The overall low predictability of the BCSE marks (14.4%) by the model may be interpreted as indicative of the influence of other variables to explain most of the variation (85.6%) in measurements. The literature review under section 5.1.5.1 refers to studies (Himmler and Schwager, 2007) (Reeves, Boyle and Christie, 2001) (Willingham, Pollack and Lewis, 2002) (Wikstrom and Wikstrom, 2005) (Thomas, Madaus, Raczek and Smees, 1998) reporting the influence of schools as a factor and the influence of student and school characteristics on the difference between tests and teachers' assessment and the variation between the two.

10. Conclusions

The conclusion attempts to summarise the major findings and the corresponding implications they have on the study, on the theory and on practice. The section is organised in order of the research questions, the results of the first question and its implications summarised in 10.1, followed by the summary and implications of the second question in 10.2. The implications of the results on the theory and on practice are presented in subsections 10.3 and 10.4 respectively.

10.1 Summary and implications of results of first question

The study appears to imply differences in the interpretation of the standards of the levels of achievement among the different measurements. The teachers' rating, student-self rating and the BCSE marks through better correlation with each other, appear to indicate the setting of higher standards relative to the teachers' award of CA marks, or the setting of lower standards by teachers in the award of CA marks relative to the BCSE marks, the student-self ratings and relative to their own rating of student competency through CA marks.

Within the interpretation of the prescribed standards of levels of achievement of the same skills of English listening and speaking competencies, the teachers and students, through their ratings, appear to impose higher standards than that indicated by teachers' assessment in the form of CA marks.

The results may be interpreted as appearing to imply a difference in teachers' interpretation of standards in their ratings of student performance and in their award of CA marks. The study (Thomas et al, 1998) explains difference between schools in variation in student scores as difference among teachers in the interpretation of the assessment criteria. However, the findings of difference within assessments (CA marks and teachers' rating) of the same teacher may also imply teachers' interpretation of standards in their assessment of student performance as being dependent on the purpose of the assessment. The influence of the purpose of the teachers' assessment, especially when high stakes are involved, is supported by the study (Hall and Harding, 2002) and predicted by the theoretical proposition of the influence of incentives on schools and teachers to behave strategically.

10.2 Summary and implications of results of second question

Within schools both the CA marks and the teachers' rating correlate better with the BCSE exam marks than with each other, while between schools findings report negative correlation of the CA marks and teachers' rating with the BCSE marks. Results show some agreement within schools but contradictions between schools with respect to CA marks and BCSE exam marks. The probable implications of these findings may be thus summarised.

Results showing low inter-rater reliability or comparability of teacher judgements between schools may be interpreted to imply differences between schools in interpretation and application of the standards

in awarding CA marks for listening and speaking competencies in English. In other words, schools interpreting standards differently explaining the relative difference in CA marks.

Results may be interpreted as being indicative of the influence of the 'adaptive level', referred to by Goldman and Hewitt (as cited in Willingham et al., 2002) and explained in the literature review as the tendency of schools and teachers to adapt the standard of their grading to student ability. The polarity in the school averages of the BCSE exam marks and the CA marks make the influence of the 'adaptive level' a probable explanation.

Thus, the implied difference between schools in the setting of standards, the better conformity of the teachers' rating with the BCSE marks than with the CA marks, the findings indicating the influence of other variables and the probability of the influence of the 'adaptive level', may be taken to indicate the probable influence of incentives as predicted by the model of the institutional effects.

10.3 Implications on theory

The theoretical proposition of the influence of incentives on schools and teachers to behave strategically when the purpose of the teachers' summative assessment has high stakes is supported by studies (Himmler and Schwager, 2007) (Willingham, Pollack and Lewis 2002) (Wikstrom and Wikstrom, 2005) discussed under subsection 5.1.5.1 of the literature review.

The outright implication of the findings as indicative of the influence of incentives offered by decentralisation cannot be established in this study of relationships. However, it may be safe to infer that the anomalies in the relationship of the CA marks with BCSE marks relative to the relationship of the teachers' and student-self ratings with the BCSE marks, are indicative of the influence of incentives offered by decentralization of high stake examinations as predicted by the model of institutional effects.

10.4 Implications on practice

Implications on practice discussed in the subsection, refer to implications on assessment practices of centralised examination systems like the Bhutan Board of Examinations (BBE).

The findings appear to imply the need to monitor the CA marks from schools in order to validate the guarantee of fair assessment of student achievement. The BBE can better justify the claim of fairness in reporting student achievement by devising methods to monitor the CA marks. Choi (1999) in referring to the importance of ensuring the credibility of school based assessments by central examination systems states,

> The Authority (The Hong Kong Examination Authority) needs an effective and efficient quality assurance and quality control system to assure the users of examination results, such as employers and tertiary institutions, as well as the general public, of the reliability of this scheme of assessment. This is not a simple task. (p. 415)

The results appear to suggest the need to standardize and moderate the CA marks, where standardization refers to processes undertaken prior to student assessment to ensure uniformity and

moderation refers to processes after student assessment to ensure consistency of grades or marks (Buchan, 1993). This translates to implications on teacher training in assessment and support to schools in the form of assessment resources. Assessment resources may be interpreted to mean uniform assessment criteria, detailed descriptors of levels of performance and standard assessment tasks. Ensuring the reliability of school based assessment used in high stakes examination systems is shown in literature to require huge investments in teacher training, assessment resources and information management systems (Harlen, 2004) (Choi, 1999) (Stiggins, 1993).

The reliability of the CA marks is best explored and monitored through longitudinal studies over cohorts and subjects. Attempts at such studies necessitate the assurance of the comparability of the examinations itself in order to facilitate objective and reliable monitoring of the CA marks. Alberts (2001) in describing the processes undertaken by the National Institute for Educational Measurement (CITO) in equating of exams in the Dutch centralised secondary examination to guarantee equivalence, highlights the importance of comparability issues in maintaining and reporting standards. Studies (Linn, 1993) (Tam, 1977) also explain the importance of the comparability of examinations and presents test linking methods to improve the comparability of examinations. When the results or marks for an examination itself is not comparable across subjects and cohorts, then studying and monitoring the reliability of the CA marks across subjects and cohorts is futile and invalid in the absence of a standard of reference that is comparable.

11. Recommendations

The chapter presets recommendations for the future studies as well as for practice in Bhutan under 11.1 and 11.2 respectively.

11.1 Recommendations for future studies

Comprehensive longitudinal studies across cohorts and subjects may serve to better address trends in school and teacher behaviour with respect to the decentralised component of centralised examinations. Comparative studies on the reliability of school-based teachers' assessment prior to and after moderation processes focusing on the differences between schools may serve to better inform reliability issues.

With regards to the issue of trust, future studies on ascertaining the trust of stakeholders (teachers, schools, parents, students, employers, higher education and training institutes) in school-based teachers' assessment is required.

A more qualitative approach, such as case studies, using observations and structured interviews with teachers may help to explore factors and conditions at the school level accounting for variations in measurement.

With regard to the standards of levels of achievement, further studies using alternative tests or observations to confirm reported level of achievement is required. Studies should pursue investigation of questions studying how well the reported standard or marks reflects what students can do.

11.2 Recommendations for practice in Bhutan

The Bhutan Board of Examinations (BBE) conducts examinations and certifies qualifications based on the curriculum prescribed by the Ministry of Education through the Curriculum and Professional Support Division (CAPSD) and the modes of assessment are developed in collaboration. Recommendations suggested are intended to inform the practices of the BBE and the support required from CAPSD to ensure the validity of its declaration of the fairness of its assessment of student achievement.

- That the BBE include in its feedback of BCSE examination results to schools, a report of performance on the continuous assessment (CA) component, instead of merely reporting the aggregated and standardised marks. Such feedback would serve to better inform schools on their assessment practices and may also serve a monitoring function.
- That the BBE study methods employed by other central examination systems of incorporating and integrating the CA component with the examination marks in the certification of the BCSE, instead of merely aggregating the CA marks and examination marks and reporting the result as a single score.

- That the division in charge of 'English' at the BBE and the CAPSD, in consultation with the English teachers, establish some standardization processes to ensure uniformity in assessment by developing a standard criteria with detailed descriptors of levels of performance and tools for the assessment of English listening and speaking skills.
- That the BBE and the CAPSD jointly explore and conduct workshops to train English teachers in the assessment of students' English listening and speaking competencies and ensure a degree of uniformity in the interpretation of the standards.
- That the BBE study methods of moderation processes that may be undertaken in a manner that is feasible, sustainable and acceptable to all stakeholders.
- That the BBE study and explore equating methods to improve the comparability of the BCSE examinations across subjects and cohorts so as to enable longitudinal studies on the reliability of CA marks and serve monitoring purposes that are both valid and objective.
- That the BBE and CAPSD study means to incorporate and support schools in the practice of student-self assessments and peer assessments in consideration of the benefits they have on the learning experience as suggested by studies (Magin and Helmore, 2001) (Patri, 2002) (Chen 2008) (Ross, 2006) as discussed in the literature review under subsection 5.2.

12. Reference list

Alberts, R.V.J, (2001). Equating exams as a prerequisite for maintaining standards: experience with Dutch centralised secondary examinations. *Assessment in Education, 8* (3), 353-367.

Bhutan Board of Examinations. (2007). *Rules and Regulations for the conduct of public examinations in Bhutan.* Bhutan: BBE

Bishop, J.H (1998). The effect of curriculum-based external exit exam systems on student achievement. *The Journal of Economic Education, 29* (2), 171 -182.

Bishop, J.H (1999). Are national exit examinations important for educational efficiency? *Swedish Economic Policy Review,* 6(2), 349-398.

Bishop, J.H and Woessman, L (April, 2002). Institutional effects in a simple model of educational production. *IZA Discussion Paper No. 484.* The Institute for the Study of Labor (IZA), Bonn.

Bishop, J.H and Woessman, L (2004). Institutional effects in a simple model of educational production. *Education Economics*, *12*(1), 17-38.

Black, P and William, D (1998). Assessment and classroom learning. *Assessment in Education,* 5(1), 7-74.

Boud, D. and Falchikov, N. (1989a). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education, 18*(5), 529-549.

Boud, D. and Falchikov, N. (1989b). Student self-assessment in higher education: a meta-analysis. *Review of Educational Research, 59*(4), 395-430.

Buchan, A. S. (1993). Policy into practice: internal assessment at 16+: standardization and moderation procedures. *Educational Research, 35*(2), 171-179.

Centre for Educational Research and Development. (2002). *The silken knot. Bhutan: CERD.*

Chen, Y. M. (2008). Learning to self-assess oral performance in English: a longitudinal case study. *Language Teaching Research, 12*(2), 235-262. doi: 10.1177/1362168807086293

Choi. C. C. (1999). Profiles of educational assessment systems worldwide: Public examinations in Hong Kong. *Assessment in Education, 6*(3), 405-417.

Curriculum and Professional Support Division. (2005). *BCSE English: teachers guide.* Thimphu: CAPSD, Department of School Education, Ministry of Education.

Charles, C.M. (1998). *Introduction to educational research*. New York: Longman.

Crooks, T. (2004). *Tensions between assessment for learning and assessments for qualifications*. Paper presented the third conference of the Association of Commonwealth Examinations and Accreditation Boards (ACEAB). Retrieved, 27th April, 2009 from: http://www.spbea.org.fj/aceab/Crooks.pdf

Drew, C. J., Hardman, M. L., & Hosp, J. L. (2008). *Designing and conducting research in education*. California; Sage publications.

Dronkers, J (1993). The precarious balance between general and vocational education in The Netherlands. *European Journal of Education, 28*(2), 197-207.

Elliot, S. N., DiPerna, J.C., Mroch., A. A., & Lang, S.C. (2004). Prevalence and patterns of academic enabling behaviours: An analysis of teachers' and students' ratings for a national sample of students: *School Psychology Review, 33* (2), 302-309.

Fuchs, T. & Wößmann, L. (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics, 32*(2), 433-464.

Gall, J.P., Gall, M.D., & Borg, W.R. (1999). *Applying educational research: A practical guide.* (4th Ed.). USA: Addison Wesley Longman, Inc.

Gross National Happiness Commission, Royal Government of Bhutan. (2009). *Tenth Five Year Plan: 2008-2013*. Bhutan: GNH Commission.

Gipps, C. (1994). *Beyond testing: towards a theory of educational assessment.* London: Falmer Press.

Hall, K., & Harding, A. (2002). Level descriptions and teacher assessment in England: towards a community of assessment practice. *Educational Research, 44*(1), 1-15. doi: 10.1080/00131880110081071

Hanushek, E. A., & Raymond, M. E. (2004). Does school accountability lead to improved student performance? *National Bureau of Economic Research (NBER) Working Paper No. 10591.* Retrieved, 3rd January, 2008 from: http://edpro.stanford.edu/Hanushek/admin/pages/files/uploads/accountability.jpam.journal.pdf

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: *Research Evidence in Education Library.* London: EPPI-Centre, social science research unit, Institute of Education, University of London. Retrieved 6th May, 2009, from: http://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=6WO5QivP0Q4%3d&tabid=119&mid=925&language=en-US

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270.

Harlen, W (2008). *Student Assessment and Testing Volume 1.* London: Sage Publications.

Harlen, W (2008). *Student Assessment and Testing Volume 3.* London: Sage Publications.

Harlen, W., & Crick, R. D. (2002). *A Systematic review of the impact of summative assessment and tests on students' motivation for learning.* Retrieved 13th November, 2008, from: http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=107&language=en-US

Himmler, O., & Schwager, R. (2007). *Double standards in educational standards-are disadvantaged students being graded more leniently?* ZEW discussion paper No. 07-016. Retrieved 5th May, 2009, from: http://www.zew.de/en/publikationen/publikation.php3?action=detail&nr=3310

HKEAA: 2010 Hong Kong Certificate of Education Examination: English language *Handbook for the School-based Assessment Component*. Retrieved on 29th December, 2009 from http://www.hkeaa.edu.hk/DocLibrary/SBA/CE-Eng-10SBAHandbook.pdf

HKEAA: *Response to concerns over SBA*. Retrieved on 29th December, 2009 from http://www.hkeaa.edu.hk/DocLibrary/SBA/About_SBA/sba_intro_c_eng.pdf

Hoge, R. D. & Coladarci, T. (1989). Teacher-based judgements of academic achievement: A review of literature. *Review of Educational Research, 5*(5), 297-313. Retrieved on 7th May, 2009 from http://www.jstor.org/stable/1170184

Kranjc, M. T. (2006). External and internal assessment in the final examination in secondary schools in Slovenia. *International Studies in Sociology of Education, 16*(2), 121-137. doi: 10.1080/09620210600849828

Kluger, A. & DeNisi, A. (1996). Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Lange, M., & Dronkers, J. (2007). How equivalent remains the central examination between schools in The Netherlands: Discrepancies between the grades of the school exam and the national central exam of secondary education between 1988 and 2005. *European University Institute (EUI) Working Paper, SPS No. 2007/03.* ISSN 1725-6755.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83-102.

Litwin, M. S. (1995). *How to measure survey reliability and validity.* USA: Sage Publications.

Magin, D. & Helmore, P. (2001). Peer and teacher assessment of oral presentation skills: how reliable are they? *Studies in Higher Education, 26*(3), 287-298. doi: 10.1080/03075070120076264

Maslowski, R., Scheerens, J., & Luyten, H. (2007). The effect of school autonomy and school internal decentralization on students' reading literacy. *School Effectiveness and School Improvement, 18*(3), 303-334. doi: 10.1080/09243450601147502

Murphy, R., & Broadfoot, P. (1995). *Effective assessment and the improvement of education – A Tribute to Desmond Nuttal*. London: The Falmer Press.

Nagy, Philip. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional Diagnosis. *Canadian Journal of Education, 25*(4), 262-279.

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing, (19)*, 109-131. doi: 10.1191/0265532202lt224oa. Retrieved on 18th May, 2009 from http://ltj.sagepub.com/cgi/content/abstract/19/2/109

Pido, S. (2005). High correlation between continuous assessment and national examination scores is achievable. *Annual Association for Educational Assessment in Africa (AEAA) Conference Papers.* Retrieved on 12th May, 2009 from http://curriculum.pgwc.gov.za/site/22/res/view/697

Policy and Planning Division. (2008). *General Statistics 2008.* Thimphu: Ministry of Education, Royal Government of Bhutan.

Powdyel, T. S. (2005). Evaluating students' achievements: The Bhutanese education assessment experience: some reflections. *Prospects, 35*(1) 45-57. Retrieved on 12th May, 2009 from http://www.springerlink.com/content/3823477577282216/

Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The relationship between teacher assessments and pupil attainments in standard test tasks at key stage 2, 1996-98. *British Educational Research Journal, 27*(2), 141-160. doi: 10.1080/01411920120037108.

Robitaille, D. F. (1977). A comparison of boys' and girls' feelings of self-confidence in arithmetic computation. *Canadian Journal of Education. 2*(2), 15-22.

Ross, J. A. (2006). The reliability, validity and utility of self-assessment. *Practical Assessment, Research and Evaluation, 11*(10), ISSN 1531-7714. Retrieved on 27th Feb, 2009 from http://pareonline.net/pdf/v11n10.pdf

Sanders, W. B., & Pinhey, T. K. (1983). *The Conduct of Social Research.* New York: CBS College Publishing.

Scheerens, J., Glas, C., & Thomas, S. M. (2007). *Educational Evaluation, Assessment, and monitoring: a systematic approach.* New York: Taylor and Francis.

Seventh Round Table Meeting, 7-9 November. (2000). *Development towards gross national happiness*. Thailand: Royal Government of Bhutan.

Stiggins, R. J. (1993). Two disciplines of educational assessment. *Measurement and Evaluation in Counseling and Development, 26*(1), 93-104.

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies, 49*(1), 26-39.

Tam, P. T. K. (1977). The moderation of internal assessments for the award of grades in public examinations in Hong Kong. *Education Journal: Hong Kong Institute of Educational Research, 6*(1), 75-91. Retrieved on 21st May, 2009 from http://www.fed.cuhk.edu.hk/en/ej/06017712.htm

Thomas, S., Madaus, G., Raczek, A. E., & Smees, R. (1998). Comparing teacher assessment and standard task results in England: the relationship between pupil characteristics and attainment. *Assessment in Education, 5*(2), 213-146. doi: 10.1080/0969594980050205

Trochim, W. M. K. (2005). *Research methods, the concise knowledge base*. Cincinnati: Atomic Dog Publishing.

Tshering, G. (2006). Student Thesis, University of Twente, *Educational Studies, 49*(1), 26-39.

Vanhoof, P., & Petegem, P. V. (2007). Matching internal and external evaluation in an era of accountability and school development: lessons from a Flemish perspective. *Studies in Educational Evaluation, 33*(2007), 101-119. doi: 10.1016/j.stueduc.2007.04.001

Vaus, D. D. A. (2002). *Surveys in social research* (5th ed.). Australia: Allen & Unwin.

Visscher, A. J., & Coe, R. (2002). *School improvement through performance feedback.* The Netherlands: Swets & Zeitlinger Publishers.

West, R., & Crighton, J. (1999). Examination reform in Central and Eastern Europe: issues and trends. *Assessment in Education, 6*(2), 271-289.

Westerheijden, D. F. (1997). Quality assessment in Dutch higher education: balancing improvement and accountability. *European Journal for Education Law and Policy, 1*(1), 81-90.

Wikstrom, C. & Wikstrom, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review, 24*(3), 309-322. doi: 10.1016/j.econedurev.2004.04.010

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: accounting for observed differences. *Journal of Educational Management, 39* (1), 1-37.

Woesmann, L. (2000). *Schooling resources, educational institutions, and student performance*: *The international evidence.* Kiel Working Paper No. 983. Duesternbrooker Weg: Kiel Institute of World Economics. doi: 10.2139/ssrn.234820. Retrieved on 11th October, 2008 from http://ssrn.com/abstract=234820

Appendix A1

TEACHER QUESTIONNAIRE
ENGLISH TEACHER ***(SECTION A)***

Dear teacher, thank you for agreeing to participate in this study.
Please fill in the blank spaces and tick the appropriate boxes.

1. Name: _____ 2. Male ☐ Female ☐

3. Name of your School : _____

4. Teaching experience: _____ years.

5. Teaching experience as English language teacher: _____ years.

6. Did you attend the orientation programme for the Revised English curriculum? Yes ☐ No ☐

Please answer the following questions with respect to 2008 when you taught English in Class 10.

7. How many students did you teach in 2008 (total number for all subjects):

around 50 ☐ around 100 ☐ around 150 ☐ around 200 ☐ around 250 ☐ over 300 ☐

8. How much time was spent in teaching/learning listening and speaking skills in the class?

never (less than 5 periods) ☐
sometimes (6 – 14) periods) ☐
Often (16 - 20 periods) ☐
Regularly (more than 20 periods) ☐

9. Which of the following activities did you have in English class?
(Panel discussion, debates, role play, reporting, dialogue)

None ☐ One to two ☐ Three to four ☐ All five ☐

10. Have records been maintained for student performance in listening and speaking? Yes ☐ No ☐

11. Did you think the guidelines for assessment of listening and speaking skills is clear? Yes ☐ No ☐

Section B requires you to rate your students against standards for listening and speaking competencies in English as specified in 'The Silken Knot'. Your ratings will be used to study the standard of English listening and speaking competencies of class 10 graduates in Bhutan. Your individual responses and ratings will be treated as confidential and will not be shared in the form of results or publications. Since honest ratings are

very important to ensure that the findings of this study are useful, I request you to rate your students judiciously against the specified standards.

Each standard has four options. Kindly rate the selected students against each standard with a [√]. You may not choose more than one option.

Please rate the following BCSE candidates of your 2008 class:

TEACHER RATING OF STUDENTS **(SECTION B)**

Name of student: _____

BCSE Index No. | **0** | **1** | **0** | **0** | **8** | | | | | | | | | | | |

1. Match their talk to the demands of different circumstances.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

2. Use vocabulary precisely and organize their talk to communicate clearly.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

3. Make significant contributions to the conversation.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

4. Evaluate the ideas and opinions of others.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

5. Take on formal and informal roles in groups.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

6. Speak in public at different kinds of functions.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

7. Use conventional patterns and forms of address in public speaking.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

8. Explain their position on and understanding of complex issues.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

9. Maintain and develop their talk purposely in a range of contexts.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

10. Make a range of contributions which show that they have listened perceptively to the development of discussion.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

11. Demonstrate an apt use of vocabulary.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

12. Participate in a variety of contexts, public or otherwise, using appropriate intonation and emphasis.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

13. Lead routine meetings and manage interactions in small groups.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

Appendix A2

STUDENT QUESTIONNAIRE
BCSE (CLASS 10) GRADUATES ***(SECTION A)***

BCSE 2008 INDEX NO: | **0** | **1** | **0** | **0** | **8** | | | | | | | | | | | |

Dear student, thank you for agreeing to participate in this study. There are no right and wrong answers. The correct answer is the 'honest' answer. *Please fill in the blank spaces and tick the appropriate boxes*.

1. Name: _____

2. Boy ☐ Girl ☐

3. Age: _____ years.

4. Name of your School : _____

5. Number of years you have been in this school: _____ years.

6. Parents occupation: farming ☐ civil service ☐ business ☐

7. What language do you speak at home? English ☐ others ☐

8. What language do you use most with friends in school? English ☐ others ☐

Please answer the following questions with respect to your time in school in 2008 when you were in Class 10.

9. You were a: day-scholar ☐ boarder ☐

10. How often did you practice listening to and speaking in English in the class?
    - Regularly (more than 20 periods) ☐
    - Often (16 - 20 periods) ☐
    - sometimes (6 – 14) periods) ☐
    - never (less than 5 periods) ☐

11. How many of the following activities did you have in English class?
    (debates, role-play, reporting, dialogue, panel-discussion)

    None ☐ One to two ☐ Three to four ☐ All five ☐

12. Do you know the standards that have been set for listening and speaking for class 10?

    ☐

Yes ☐ No

13. Did you know how much you scored for listening and speaking (20% CA?)
Yes ☐ No ☐

STUDENT SEFL-RATING ***(SECTION B)***

This section requires you to rate yourself on different aspects of listening and speaking skills in English. Your ratings will be used to study the standard of English listening and speaking competencies of class 10 graduates in Bhutan. Your individual responses and ratings will be treated as confidential. They will not be shared or published in the form of individual results. Since honest ratings are very important to ensure that the findings of this study are useful, I request you to rate yourself truthfully.

The following section has questions on different areas of listening and speaking competencies. Each area has four options. Kindly rate yourself in each area with a [√]. You may not choose more than one option.

1. When a person speaks to me ***in English***, I am able to continue the conversation ***in English***.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
2. In a class debate or extempore speech, I am able to speak clearly and fluently to express my ideas ***in English***.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
3. When speaking to teachers and elders ***in English***, I am able to use suitable English words and sentences to show respect.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
4. When asked a question in class which needs to be answered in length ***in English***, I am able to use different words to answer effectively.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
5. During a debate ***in English***, I am able to listen to and understand the speakers.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
6. During a debate ***in English***, I am able to counter the opponents with questions and comments.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
7. While conversing with friends ***in English*** during group work in class, I am able to understand the group discussions.
   Not True At All ☐ Not True ☐ True ☐ Very True ☐
8. While conversing with friends ***in English*** during group work in class, my ideas and suggestions are noted and appreciated.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

9. While listening to an interview ***in English*** on television or the radio, I am able to understand the conversation.

   Not True At All ☐ Not True ☐ True ☐ Very True ☐

10. While listening to an interview ***in English*** on television or the radio, I am able to agree or disagree with the opinions and ideas being discussed.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

11. When a very good speaker is delivering a speech ***in English*** in the morning assembly, I am able to understand the speaker.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

12. When a very good speaker is delivering a speech ***in English*** in the morning assembly, I am able to agree or disagree with the speaker's ideas and opinions.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

13. When planning the celebration of Teachers Day in class, I am able to communicate effectively ***in English*** with my class mates to organize the programme.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

14. During the farewell of a teacher trainee from NIE, I am able to deliver a vote of thanks ***in English*** on behalf of the class.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

15. During a class/school debate, I am able to speak effectively ***in English***.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

16. When I have to give a presentation of my work in class, I am able to deliver it clearly ***in English***.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

17. When delivering a formal speech in class, I am able to acknowledge the audience and introduce my topic ***in English***.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

18. During a debate in class, I am able to take on the role of the chairperson and properly introduce the topic, the groups and explain the rules ***in English***.

    Not True At All ☐ Not True ☐ True ☐ Very True ☐

19. A poem is being discussed in class for a lesson in English. When asked by my teacher, I am able to identify the theme of the poem and justify my point ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

20. When presented with two solutions to a problem by my teacher, I am able to make a choice and support my choice ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

21. I have missed a class because I got up late. I am able to explain to my disappointed teacher the reason for my absence and apologize effectively ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

22. I am helping a cousin from the village at the hospital. I am able to explain to the doctor how she feels ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

23. The teacher has asked the group to discuss and present the conclusion of 'The Giver'. During the group discussion ***in English***, there are many opinions and justifications. As group leader, I am able to effectively summarize the different views for the presentation.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

24. When two of my friends are having an argument and I need to help them, I am able to give suggestions ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

25. When talking ***in English*** about a person or an event with my friends, I am able to use different adjectives (describing words) in my description.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

26. When trying to express my feelings and emotions with close friends I am able to use words that explain exactly how I feel ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

27. During an extempore speech, I have been given a topic that I like. I am able to express my surprise and good luck for getting the topic in my speech ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

28. I have been voted class captain by my class mates. When asked to say a few words, I am able to express my gratitude and hopes effectively to the class ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

29. The class is planning a programme for 'Teacher's Day'. I am able to conduct the meeting ***in English*** and develop the programme with my classmates.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

30. My group has been asked to prepare and present a sample letter of invitation. As group leader, I am able to organize the task and manage the discussions ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

31. I enjoy listening to and speaking ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

STUDENT SELF-RATING ***(SECTION C)***

This section requires you to rate yourself on writing, language and reading and literature. Your individual responses and ratings will be treated as confidential. Since honest ratings are very important to ensure that the findings of this study are useful, I request you to rate yourself truthfully.

The following section has questions on areas of writing, language and reading and literature competencies. Each area has four options. Kindly rate yourself in each area with a [√]. You may not choose more than one option.

32. While writing an essay ***in English*** during a class test I am able to spell correctly almost all the words.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

33. If I wrote an essay ***in English***, it could be selected to be published in the school magazine.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

34. My aunty wishes to apply to the Bank of Bhutan for a loan to start a business. I am able to confidently write an application ***in English*** to the bank for her.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

35. When my essay is returned by the teacher, it hardly has any red ink to show my grammatical errors.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

36. I often use appropriate idioms and proverbs when I speak and write ***in English***.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

37. I am generally confident that my spoken ***English*** is grammatically correct.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

38. If asked, "What kind of literature do you like to read and why?" I have an answer.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

39. When I come across a new poem in a book or newspaper, I am able to read and understand it.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

40. I can say that I read a lot in addition to the literature in the school texts.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

STUDENT SEFL-RATING ***(Section D)***

This section comprises questions on academic self-concept taken from *OECD's Brief Self-Report Measure Of Educational Psychology's Most Useful Affective Constructs: Cross-Cultural, Psychometric Comparisons Across 25 Countries* by Marsh, Hau, Artelt, Baumert and Peschar (2006).

Kindly rate yourself in each area with a [√]. You may not choose more than one option.

41. I learn things quickly in most school subjects.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

42. I am good at most school subjects.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

43. I do well in tests in most school subjects.

Not True At All ☐ Not True ☐ True ☐ Very True ☐

Thank you for your participation and your time.

Appendix A3: Outputs of Rasch Analysis of pilot items of questionnaires using "Quest"

```
TEACHER QUESTIONNAIRE Y08
------------------------------------------------------------------------
Item Analysis Results for Observed Responses
16/12/2008 21:55
all on all (N = 45 L = 12 Probability Level=0.50)
------------------------------------------------------------------------
Mean test score       19.18
Standard deviation     7.71
Internal Consistency   0.96

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is a limited amount of missing data.
========================================================================



STUDENT QUESTIONNAIRE Y08
------------------------------------------------------------------------
Item Analysis Results for Observed Responses
16/12/2008 22:31
all on all (N = 45 L = 46 Probability Level=0.50)
------------------------------------------------------------------------
Mean test score       49.40
Standard deviation    10.79
Internal Consistency   0.89

The individual item statistics are calculated
using all available data.

The overall mean, standard deviation and internal
consistency indices assume that missing responses
are incorrect.  They should only be considered useful when
there is a limited amount of missing data.
========================================================================
```

Appendix A4: Letter of approval for data collection from schools

ROYAL GOVERNMENT OF BHUTAN
MINISTRY OF EDUCATION
DEPARTMENT OF SCHOOL EDUCATION
THIMPHU : BHUTAN

Ref. No. MoE/DSE-1/2009/ 2969 6th January 2009

To Whom IT May Concern

This is to convey the approval for Dechen Dolkar, Bhutan Board of Examinations to engage in a research to ascertain the standard of English listening and speaking skills of class 10 BCSE 2008 graduates. This research is conducted as a partial fulfillment for her Master Study Programme in Educational Science and Technology at the University of Twente, Netherland. The research would require her to visit some of the Higher Secondary Schools as listed below:

- Punakha HSS.
- Bajo HSS
- Mongar HSS
- Gyelpozhing HSS
- Jigmesherubling HSS
- Gelephu HSS
- Damphu HSS
- P/ling HSS
- Drukgyel HSS
- Yangchenphu HSS
- Kelki HSS
- Ugyen Academy HSS

Concerned Principals and DEOs are requested to give her all the support.

(Tshewang Tandin)
DIRECTOR

Tele: 00975-2-325325, Telefax: 325141

Appendix A5: Instructions for administration of questionnaire

Dear _____

Thank you for accepting to help in this study.
Below I have some guidelines that will help to ensure proper data collection.

1. I would first require to get a list of all the BCSE 2008 candidates of your school who are now in your school in class 11. I require only their index numbers to be indicated in the same column under the name of the English teacher who taught them in 2008 when they were in class 10. This is required so that I can randomly select 15 students for each of the 2008 class 10 English teachers.
2. The list of students (with the BCSE index numbers)to be sent to fax number 02 3223290. I will then generate a random sample of 15 students against each English teacher.
3. The questionnaires for teachers are meant only for those teachers who taught English in class 10 in 2008. The questionnaire requires them to rate 15 of their 2008 class 10 students. The 15 students to be rated are the same 15 students who are selected for the study. Thus, each teacher who taught English in 2008 will rate the 15 students selected by me from the list given.
4. The student questionnaire will be administered to those 15 students selected. They are requested to complete the questionnaire and rate themselves.
5. Number 3 and 4 mean that the same students who are selected for the study will complete the student questionnaire and these same 15 students will be rated by their 2008 English teacher.
6. Teachers and students should work on their own in completing the questionnaire. Teachers may be given up to 2 and half hours to complete the questionnaire and ratings of 15 students and students may be given 15 minutes. I request that teachers do it in school within the stipulated time and in a quiet room. All teachers who are part of the study may do it in the same room. Students may kindly be asked to complete the questionnaire quietly within 15 minutes in a room.
7. Kindly explain to the teachers and students that this data is for the study only ***and no individual reports for teachers and students will be compiled. Teachers and students participating will not get any individual report and rating.*** The purpose of the study is mainly to get a measure of the BCSE 2008 graduates standard of listening and speaking skills from the perception of students and teachers. Honest answers would really improve the validity and sensibility of the study. Kindly stress the need for ***honest*** answers and ratings from the students as sometimes they think that they will be rewarded or penalized. Honesty is the greatest contribution they can make to this study otherwise all effort, time and energy is in vain.
8. Kindly thank all the teachers and students for taking part in the study on my behalf and until such time I can come personally to meet and thank them.
9. The completed forms be kindly collected after ensuring that all questions have been answered.
10. Kindly arrange to send the documents to me. Costs for delivering the documents be kindly billed to me.
11. Address for documents and bills to be sent to either of the two addresses:

Dechen Dolkar Bhutan Board of Examinations Ministry of Education Thimphu Bhutan	Dechen Dolkar c/o MB Ghaley Save the Children Post Box no. 281 Thimphu

12. Last but not the least, my sincerest gratitude for your assistance and help.
13. If you have any clarifications, kindly give me a miss call at 17604364 or a short call at 02 336221 so that I can call back.

Gratefully Yours,

Dechen Dolkar

Lists of teachers and students participating in the study:

(Kindly provide the names and BCSE 2008 index numbers of students taught by the respective teachers who passed the BCSE 2008 examinations and are currently in class 11 in your school. Kindly indicate the index numbers of students under the name of the English teacher who taught them)

Name of Teacher……..		Name of Teacher……..	
Sl. no	Student BCSE Index no.	Sl. no	Student BCSE Index no.

(Kindly use additional paper if required.)