Non Standardized language Tests, Faulty evaluations & it's impact on Pakistani educational setup

An M. Phil (ELT) Dissertation

submitted To:

The Department of ELT, Kinnaird College

By

Ayesha Aziz

Acknowledgements

I would first of all like to thank Dr. Asma Aftab, my research supervisor, for her tireless support and encouragement throughout the drafting of this thesis. I would also like to express my heartfelt thanks to Dr. Tajjamil Hussain, Assistant professor in the Department of Mathematics at COMSATS, University, for his patient and unflagging help with this research thesis. Without the professional help and devotion of him, the technical part of this research would not have been compiled.

I would like to thank Professor Dr. Ahmed Kaleem for his help with this research thesis. His valuable comments were invaluable in drafting several chapters.
I would also like to thank all of my colleagues, who helped in the data collection to conduct this research.

I am extremely grateful to Jayanti Banerji and Carolina Clapham from The University of Lancaster(UK) for their unfaltering support over my entire research.
I am also grateful to the Journal management of "Francis & Tylor" who contributed to this research by guiding me through their valuable recommendations about the innovative concept of benchmarking and rater training in Pakistan.

Abstract:

The examination mechanism in Pakistan has many loopholes, the local benchmarking is entirely not existing, many Pakistani examiners are not trained in the application of rating tests accordingly the scale descriptors, due to this students do not get vivid feedback about their test performances. Therefore, this study is designed to facilitate and initiate the interest of the English language teachers in Pakistan to understand the importance of sound assessment and validation mechanisms, and their exposure regarding the construction of test items to assess the English language proficiency of Pakistani students at both the summative and formative levels. This study is carefully selected and proposed while keeping the most erroneous practices in Pakistani examination systems. Most of the English language teachers in Pakistan are entirely unfamiliar with scaling and testing procedures which play a crucial role in setting the pupil's language proficiency targets. As a matter of fact, the Pakistani students lag behind in International tests, because they are not tuned to the highly mechanized examination systems.

The primary focus of this study is to assess the test validity through the construct validity; content validity and criterion related validity of the test items. The obscurities faced by Pakistani students in acquiring English language proficiency are also assessed by the summative evaluation through local assessment screening tests likewise: NAT1&, NAT2. This study has also provided with sound recommendations in improving pupils' proficiency in such local assessments and it also highlights test takers anticipated fallacies and fears regarding their test performances.

Thus, this study emphasized the sound interrelation of the summative and formative benchmarking both at the institutional level as well as governmental level.

Abbreviations:

National testing systems	NTS
Graduate Record Examination	GRE
Graduate Management Admission Test	GMAT
Test of English as a Foreign Language	TOEFL
International English Language Testing System	IELTS
National Admission Test	NAT
Graduate Assessment Test	GAT
United Nations Educational, Scientific, and Cultural Organization	UNESCO
International Association for Educational Assessment.	IAEA
International Development Program of Australian	IDP
Universities and Colleges Ltd	UCL
Non-standardized Evaluation Practices	NEP
Formative/Summative Assessments	FSM
Inefficacy of National Testing Systems	INT
Non unified Assessment Criteria	NAC
International Computer Technology	ICT
Computer Based Test	CBT
Multiple choice Questions	MCQS
Statistical Package for Social Sciences	SPSS

TABLE OF CONTENTS:

LIST OF TABLES:

Chapter1: Introduction

This study focuses primarily on the role of local testing bodies in the Pakistani examination system. and these constitute the corner stone of Pakistani educational setup. This research examines the intricate educational policy issues. Pakistan since it's inception is facing numerous problems in defining clear educational policies. One of the major reason is the political upheaval due to which many educational reforms remained in the pipeline. The dual mediums of instruction - Urdu and English medium - has only created a chaos and extreme dejection amongst less privileged classes which can not bear the expenditures of standard education; on the contrary, the privileged class laments about bearing the heavy expenditures of education but still lags behind in competing the international standards of education.

The local testing bodies do not pay adequate attention on the deteriorating examination standards. A unified assessment system may be required; there is a constant debate going on whether there should be unified testing body or the universities must be given full authority to take their own aptitude test. This study examines the situation through a survey focusing on the following research questions.

1.0 Research Hypothesis:

This research has following research question i.e. **"Why the non standardization is existing in Pakistani testing and evaluation system"?**

To address the above stated problem, four research questions were formulated: the main aim was to understand the root cause of non standardization in Pakistani examination systems. These research questions are:

- What are the key factors behind non standardization which is affecting the Pakistani examination system?
- What are the reasons behind non descriptive and irrational marking related to both summative and formative assessments in Pakistani educational setup?
- Why does the Pakistani local testing bodies remain ineffective in implementing standardized testing in local examination systems?
- What are the factors affecting local assessment criteria in the Pakistani examination systems?

According to one Pakistani educationist Dr. Jilani Warsi "almost all tests in English in Pakistan, are, in fact, examinations: they are subjective in their setting and marking in that they only cover the skills of reading and writing. They measure the pupils' knowledge of the language rather than their performance in it, and they confuse the testing of language with literary and cultural attitudes and knowledge. It goes without saying that formal tests are an attempt to construct an instrument for measuring ultimate attainment, or progress, or ability in language skills.

Most English language exams exhibit a glaring omission of the below mentioned objectives. It is needless to say that the learner stands to gain from these objectives, as knowing about implicit cultural values of the target language expands her or his world view.

The objectives of second language pedagogy and assessment should be:

- To develop the learner's intellectual power through a second language, in this case English.
- To enhance the learner's personal cultural values through the study of second Language, literature & philosophy.
- To increase the awareness of the mechanism of the learner's native language through a second language.
- To keep the learner abreast of current writing and research in his/her respective discipline.
- To enable students to communicate orally in their second language.
- To assist the learner in acquiring reading, writing, listening, and speaking skills.
- To expand gradually deepening knowledge of a foreign country.
- To help the learner appreciate cultural experiences through improved second language skills.

On the contrary, the Pakistani English language teachers seem to be less motivated in understanding the challenging but interesting mechanisms of developing language proficiency assessment criteria to equip their students to compete effectively with the challenges of the higher levels of education in Pakistan. This study aims to identify reasons behind this phenomenon; the anticipated in this research:

There is little or no research focusing on English language Testing and proficiency measurement in Pakistan. The Pakistani teachers are not adequately trained to understand the basic phenomena of content, construct & criterion validity. The trialing procedures of the test items are not in fashion in Pakistan.

The Pakistani English language teachers are not accustomed to the International Testing &Scaling procedures. Little guidance and funding support discourages the English language teachers to conduct pilot projects on language proficiency tests.

Before probing into the above stated issues, it is essential to highlight the common fallacies which raise misconceptions about various scales. Basically, scales are not tests: they may "provide a means of describing levels of proficiency,[but] do not in themselves measure that proficiency." (Rudd: 140). This means that for assessment purposes, the scales need to be used in conjunction with a language test or task; the test measures language ability and the scales describe language ability as measured by the test. This study also proposes to understand the key jargons and to understand the definitions of the test's content validity, construct validity, & criterion related Validity.

The theme of this research is standardization in testing. One reason for selecting this theme is this there are certain explicit criterias for construction, administration, analysis and evaluation of tests that are not incorporated in the Pakistani examination system, and this scenario is badly affecting the English language testing in Pakistan.

There are many aspects related to standardization in testing in the Pakistani educational set up: it refers to the procedures and terminology for describing test contents, what it intends to measure, its format, administrative procedures and so on
The American Psychological Association has issued various publications to stimulate standardization of testing practices such as the "Standards" (1974) and the more recent "Code" (1988). This article focuses on two important qualities of tests viz., reliability and validity and standardization of procedures and techniques to secure and demonstrate them.

1.1 Justification of this study:
It is assumed that the above stated issues require special attention because tests are the ways of controlling and imposing specific knowledge of students, teachers, principles and educational systems alongside certain agendas and educational ideologies (Broadfoot, 1996; Shohamy,2001).

This study emphasizes establishing the quality of measurements consisting of various stages, each related to a particular aspect. To begin with, there is the aspect of reliability i.e. the accuracy or stability over time with which a trait (attribute, skill, knowledge, or insight) has been measured. It is believed that Pakistani local examinations do not pay attention to this essential aspect. Reliability concerns the question whether differences in scores between subjects reflect differences in the attributes measured or whether they are the result of
Accidental factors within the test (which may be badly constructed or may not incorporate sufficient items/questions) or without(cheating, underachievement due to indisposition etc.
If accidental factors have unduly influenced the test results, administration of an equivalent test at the same time or of the same test at a different time would have yielded a different rank order of the subjects; in other words: the scores cannot be interpreted as an adequate reflection of a trait.

1.2 Limitations of this research: This research has certain limitations, it is restricted to one of the cosmopolitan city of Pakistan (Lahore) in the province of Punjab in Pakistan; the research data was not collected from the different colleges/universities of the entire country. Due to its demographic limitations, the results of this study cannot be generalized either to the other Pakistani examinations or the world wide examination systems, The population involved in this study were limited to one hundred students and forty English language teachers from the two public and two private sector universities of Lahore. It is also limited in its nature because it has examined only a single local examination body that is NTS,(National testing systems) therefore, the research findings of this study are restricted to the local examination systems, and the data implications of this study are unknown in case of other examination systems both in Pakistan as well as across the world.

1.3 Structure of the Dissertation:

Chapter 2: provides the conclusive literature review comprising the assumptions of the various researchers about the importance of standardized tests, educational standards and problems of error while rating the tests. It also included exhaustive but constructive criticism on the National testing systems in Pakistan.

Chapter 3: thoroughly discusses the methodology adopted in the present research. It focuses on the sampling procedures, method of data collection, and data analysis which can help exploring the defined research problem. The mixed method research approach was applied to achieve the results. Percentiles and correlations were utilized to highlight the respondents' opinion. Chapter 3 further provides detailed discussion on the framework used to develop research relevance with the current testing and evaluation issues both at the formative and summative levels.

Chapter 4: comprised of the detailed summary of survey results and it also highlights the correlation of the research objectives

Chapter 5: comprised of the conclusion and recommendations which shows the significance of this research. It also consists of a detailed discussion on the strong link of the four research variables. The recommendations stated in this chapter are assumed to be effective in devising a futuristic approach towards the sound examination mechanisms in Pakistan.

Chapter 2:

2.0 Literature Review:

The Pakistani examination system is one of the corner stone in the educational setup of Pakistan, this study focuses on multiple discrepancies which are becoming a major obstacle in developing a transparent and unified examination system in Pakistan at the tertiary level. Therefore, both the students and their parents feel extreme anxiety and dejection even after bearing heavy expenditures on education. According to Trochim: "The generic goal" of most evaluations is to provide "useful feedback" to different audiences and that evaluation should influence feedback". Whereas, the Pakistani examination system has failed to provide sufficient feedback due to dual educational and curriculum design policies, there are insufficient local benchmarks to evaluate students in a uniformed manner. Educational colleges and institutions do not rely on the local testing bodies, thus this scenario reflects lack of standardization which is the key requirement of advance standards of education in any society.

This study is conducted to highlight the errors in measurement, which make the entire Pakistani examination system as insignificant and unnecessarily imposed procedures. It is diagnostic in it's nature and is designed to probe into the answer to a fundamental question: i.e. How is error in measurement of standards obscured the assessment of candidates both at summative & formative levels? Therefore, it is essential to know what exactly an assessment or examination is? According to UNESCO (1961) "Examination is a measuring instrument, intended to verify both candidates value and value of the teaching he has received. It is an indicator of training given and received. It is to measure what has been accomplished during the period of study. To weigh each candidate's sum of knowledge and appraise his ability. It looks like target, incentive, motive or stimulant. "It provides motives for the teacher and a spur for the pupils. Examination conditions orientates the entire teaching process."
Barnard & Lauwerys (1967) explain the concept of examination as "A test of knowledge acquired, or more generally a means of assessing intellectual capacity or ability. According to them, there are normally three types of examination:

(1) a set of questions designed to check pupils progress on the results of a course of instruction.
(2) a means of qualifying candidates for a certificate or degree in which they are required to attain a certain standard for a pass or honors.
(3) a competitive test on the strength of which a scholarship or other award is made to successful candidate. Examination may be conducted by means of written answers to set questions or by local methods."

Page & Thomas (1978) explain the concept of examination as (1) Assessment of ability, achievement or performance in a subject. (2) Instrument of assessment can be log essay or mixed form of assessment. May be used for qualifying for entrance to professions and higher education

Rowtree (1981) explains the concept of examination in these words, "A formal assessment of student's learning, used particularly at the end of the course. Although many teachers bemoan the distorting influence examination can have on teaching and learning. When used wisely they can assess students' qualities that would otherwise go unobserved. An examination usually involves one or more of the following features: All students are given the same task to perform (perhaps with some degree of choice); it's precise nature; will not be made clear

until the moment they begin upon it; they will be given a time limit and they will not be allowed to consult references or one another; they will perform in the presence of invigilator; and they will expected to experience some sense of urgency or stress. Now-a-days, many examination one or more restrictions (e.g. students may be given advance notice of the questions or may be allowed to use certain reference materials, but the last mentioned feature usually remained" (Aggarwal.1997).

2.1 What are Formative & Summative Assessments:

In order to explore the nature of the term, consider two recent and prominent definitions. Popham (2008) defined formative assessment as:

… a planned process in which assessment-elicited evidence of student's status is used by teachers to adjust their ongoing instructional procedures or by students to adjust their current learning tactics .

The Council of Chief State School Officers (CCSSO) defined it as:

… a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructionaloutcomes (CCSSO, 2006 cited in McManus,2008.)

Note that in both of these definitions, the focus is on the process or set of actions and not on the assessment objects themselves. Implicit in these definitions is the assertion that in order to have formative value, an assessment must be done – and the results used – within a process that occurs during an instructional unit, provides accurate and relevant information about student performance, and is coupled with various strategies to generate information as to where the student is now and where to go next with instruction. We need to be explicit that in a system with formative value, what goes on around the time the student takes the assessment is as important as what goes on during the assessment. Such a system could be described in a number of ways, however for the purpose here, content, context, and strategies are highlighted as necessary components as a means to suggest a better way of phrasing formative processes.

Labeling the item itself as inherently formative or summative ignores important and necessary considerations related to the item such as timing, alignment to instruction, and what the student and teacher do with the information obtained. As Wiliam (2000) noted: It has become conventional to describe these two kinds of assessment as formative and summative assessment respectively, but it is important to note in this context that the terms 'formative' and 'summative' do not describe assessments – the same assessment might be used both formatively and summatively – but rather are descriptions of the use to which information arising from the assessment is put.

Teachers/ Examiners are not properly trained in understanding the assessment mechanism, they usually copy the test items from somewhere, used to conduct an exam and then evaluate it either in terms of Marks or grades, the faulty educational practices and lack of teachers own interest never allowed them to think that "what exactly they are delivering to their students & in terms of what manner they have to monitor the output, in fact, most of them are not interested in the output.

Another reason is the unwisely selection of the curriculum, this research probed into the thorough survey of syllabuses and it is observed that either the contents of text books are too easy to be taken for granted or too distinct to be drab because the social phenomena's are borrowed from foreign cultures.

Dual medium of instruction likewise Urdu/ English mediums is the another major reason, which is the major obstacle in the improved and standardized local examinations in Pakistan,

because the unified curriculum policies are not in real implemented, therefore, our local examination system is bound to make the tests too simplistic and at times substandard.

This research has also highlighted the assessment mechanisms, the band or proficiency descriptors are not in use, if you ask an examiner that what exactly the criteria upon which you have given marks either high or low to the candidate? The reply is not substantial because every examiner is having his/her criteria to assess rather than following a standardized evaluation practice, because marking descriptors or bands are not in practice, as a matter of fact errors in measurement are obvious but on the other hand it has a disastrous impact because Pakistani students face great difficulty in competing in International exams likewise: GMAT, GRE, TOEFL & IELTS. This research probes in to the depth of what exactly is the key criteria upon which a Pakistani evaluation system must be built on, in other words one must say that "it is the evaluation of Pakistani testing & Evaluation systems in itself".

It is essential to discuss the need for assessments, many of the researchers emphasize that the assessment tasks must be real life tasks. According to Bachman (2002) "a fundamental aim of most language performance assessments is to present test takers with the task that correspond to tasks in real-world. Moreover, Eisner (1999) emphasized on the performance based test tasks, he described "performance assessment" either the observation of behavior in the real world or a simulation of a real-life activity with raters to evaluate the performance. Neol Wilson's one of the article "Educational Standards and the Problem of Error". Where he rightly stated that: "In the myth of meritocracy the examination is both a major ritual and a significant determinant of success. At the heart of this ritual, between the practice and the judgment, between the stress and the catharsis, is the great silence, the space where the judgment is processed. The myth gives hints of what moves in this silence, for the myth makes three claims: the race is to the swiftest; the judgment is utterly accurate; and success is a certification of competency.

These hints tap the bases of the three frames of reference for assessment that assume objectivity. However, other assumptions of these frames make them mutually contradictory. This in itself would be good reason for keeping the process implicit. For the assumption that inside the black box hidden in the silence is a mechanism, an instrument of great precision, may be difficult to sustain, if it contains majorcontradictions within its workings".

According to Neol: "Four assessment systems, with four different frames of reference, have staked their claim to exclusive use of the black box, their claim to be the best foundation for the precision instrument to measure human - what? Bit hard to say what exactly. To measure, perhaps, human anything. It may be sufficient just to measure. Or even just to pretend to measure, to assert that a measurement has been made, so that a mark may be assigned to a person.

2.2The Judge's frame: is far more often evoked than talked about. The focus is on the assessor's judgment of the product. The major activity is in the mind of the assessor. Such terms as expert and connoisseur are essential to the construction of the accompanying myth. Faith is the requirement of all participants. It is explicit in discourses about teacher tests, public examinations, and tertiary assessment, and implicit in all human activities that involve the categorization of people by assessors.

2.3 The General frame: is the basis for educational measurement, for psychometrics. The focus is on the test itself, its content and the measurement it makes. Such terms as reliability and ability are essential to its mythological credibility. It purports to be objective science, and

hence independent of faith. As such the world it relates to is static, so there is no essential activity. It is explicit in discourses about educational measurement, standardized tests, grades, norms; it is implicit in most discourses about standards and their definitions.

2.4 The Specific frame: is about the whole assessment event, and is the basis for the literature that derived from the notion of specific behavioral objectives. The focus is on the student behavior described within controlled events; in these events the context, task, and criteria for adequate performance are unambiguously pre-determined. Reality is observable in the phenomenological world; the essential activity is what the student does. This frame is explicit in discourses about objectives and outcomes; it is implicit, though rarely empirically present, in discourses about criteria, performance, competence and absolute standards.

2.5 The Responsive frame: focuses on the assessor's response to the assessment product. Unlike the other frames it makes no claims to objectivity; as such its mythical tone is ephemeral, its status low. This frame is explicit in discourses about formative assessment, teacher feedback, qualitative assessment; it is implicit though hidden in the discourses within other frames, recognized by absences in logic and stressful silences in reflexive thought. Within the confines of communal safety such discourses are alluded to, skirted around, or at times discussed; on rare occasions such discourses emerge triumphantly as ideologies within discourse communities.

2.6 The Judge:

Most assessment in education is carried out within the Judge's frame of reference. The chief characteristic is that one person assesses the quality of another person's performance, and this assessment is final. By definition the Judge's assessment is free of error, and therefore any check of the Judge's accuracy would represent a contradiction of his function. So such a check is not only unnecessary, it is immoral, in that it is an act likely to destabilize the whole assessment structure by calling into question its most hallowed assumption. The Judge's assessment may be verbal and on-site, eschewing numeration and a special testing context. However, performance is usually assessed with tests and examinations, with merit graded in some way. It is assumed that adequacy or excellence in performance is described accurately by the Judge. For this to be true, it must also be assumed that the test measures what it purports to measure, and that the marking, whether by the Judge or his assistants, is reliable. Again, therefore, checks of validity, that the test measures what it purports to measure, or of reliability, that the test will give the same result if repeated, are not only unnecessary, but are unacceptable and demeaning. Judges must stand firm on the absoluteness and infallibility of their judgments, for this is the essence of their power, the linchpin of their role, the irreducible minimum of their function.

Thus they are duty bound to recognize standards, to perceive with unerring eye that thinnest of lines that separates the good from the bad, the guilty from the innocent, the excellent from the mediocre, the pass from the fail. Talk to them of normative curves or rank orders or percentiles, all of which imply relative standards, and they will hear you out, wish you well, and with scarcely disguised distain send you on your way. In their absolute world such matters are irrelevant. They know what the standard is, and therefore their job is simple. Simply to allocate students, or their work, to various positions above or below that standard. Set hard in a rationalist world view, this is a black and white world, a fundamentalist cognitive universe. The assumptions deny the possibility of reality checks, so the collective fantasy easily becomes the perceived truth, as human minds and bodies contort themselves to deny their more immediate experience. So let us see what that more immediate experience might tell us if another frame of reference is chosen.

2.7 The General:

The second frame of reference is called the General frame. Neol used to call it the generalizability frame, but that word has been hijacked by psychometricians. The general has been privatized and corporatized by mathematicians. The bird has been tamed and lost its wings. The general has become severely contained in mathematical armor. What Neol is calling the General frame of reference is blatantly egalitarian and inherently relativistic in its conception, but has become constricting, reductionist and inequitable in its mathematical application. In one form or another it has dominated the academic literature in educational assessment for over sixty years. Within this frame is contained most of the received wisdom from thousands of studies in educational measurement and evaluation.

Its two initial assumptions are shattering. One Judge is as good as another. And all Judges are inaccurate. On the contrary to the above said criteria's of Neol Wilson, our examination system does not befit on even a single frame, most of the test stuff is plagiarized, unknowingly the logic of adding a specific item.

In the above stated preview of the researchers expectations, this study is going to throw light on the local assessment bodies and summative/formative assessment criteria which is unsatisfactory and beyond the realistic academic demands of Pakistani students. In the review of literature of this study, it is essential to discuss National testing systems in Pakistan.

2.8 National Testing Service (NTS) is an organization in Pakistan that conducts academic performance evaluation tests. It is similar to Educational Testing Service (ETS) in the United States. NTS offers two main types of tests, the National Admission Test (NAT) and the Graduate Assessment Test (GAT). NAT is aimed at students seeking admission to colleges and universities. GAT is aimed at graduates seeking admission to postgraduate education. NTS exams are also used to determine qualifications of students seeking advanced study abroad.

NTS is a Member of the International Association for Educational Assessment, USA (IAEA). It is recognized by the Higher Education Commission of Pakistan (HEC). The NTS was developed to ensure quality educational standards in Pakistan and to "provide a national scale for comparative grading between institutes", consolidating examination boards under one administrating body. According to Shahid Siddiqui, director of The Centre for Humanities and Social Sciences at the Lahore School of Economics, tests implemented prior to the development of the NTS were criticized as not accommodating socio-cultural differences, resulting in a need for "an indigenous testing service that should design and develop testing materials within an indigenous context".

2.9 History of NTS:

Controversy erupted in 2007 following the mandating of NTS testing, first announced by Khalid Maqbool, Governor of Punjab, with regards to admission to universities and later set aside by the government unless the laws regarding university admission were amended. Criticism included allegations that the services were inconvenient to access and prohibitively expensive to economically challenged students. Maqbool called for input from the vice-chancellors of Punjab's public universities on 4 June. That year, each university department was permitted to set its own entrance requirements, with six of 64 departments at PunjabUniversity electing to utilize tests from NTS. In October 2007, the vice-chancellor of Punjab University, Muhammad Arif Butt, endorsed the use of NTS testing in all departments of the university. NTS director Dr. Haroon Rasheed said that no tuition centres are associated with the service and NTS exams will not be out of course.

2.8 Criticism About NTS:

NTS has sought to establish a national educational standard for Pakistan but at the same time as an organization cannot and has not catered to the regionally diverse Pakistani nation. NTS resources are geographically restricted and both quantitatively and qualitatively limited. As stated by interface - an educational consultancy. Accessed Mar 16, 2010.</ref> Its contention as an independent NGO can also be challenged as it is affiliated with numerous regular testing centers as semi-Government body or inter-Governmental organization & Public / Governmental education Institutions for their entrance / admission exams. A recent controversy of regular operational error has also arisen for Phd. assessment tests. The credibility / irresponsibility on the part of NTS officials has disqualified the candidates from applying in MPhil/PhD programs at public and private sector institutes because under HEC rules, they do not have any proof of clearing the test. The overall process lacks credibility & transparency due to the un-disclosed nature of alternative processes for any participants, when they do not fall in the norms of the operating procedure (i.e. errors / loop holes) due to complexities that arise from multifarious of socioeconomic nature of challenges in present day environment of Pakistan. Unfortunately, the NTS management have even been caught and identified as plagiarizing test questions and are indirectly controlled by (HEC).

After having a brief snapshot of the functionality of the Pakistani local examining body, this study leads to various factors which are required to be implemented to make the Pakistani examination system as more transparent and viable, these are: The Assessment Authenticity must be assured through the frequent validation of the testing criteria .Process, Performance, Products and Portfolios need to be considered and no compromise will be made in case of it .Authentic Assessment in subject areas require serious attention in Pakistani context.(Massive surveys can be conducted & educationist must directly be involved in to the test item validation procedures.

Cultivating the sense of self assessment among candidates, so the demotivation must be reduced among the masses. Many studies reveal that the summative and formative assessments go side by side. No matter to what extent the Pakistani policy makers try to bring reforms to the examination systems this cannot be improved, unless the classroom teaching is not going to be standardized. Unfortunately the classroom teaching practices are not up to the standards of quality education in Pakistan.

A well accepted position among educational researchers and teacher educators is that the *best* classroom assessments are *authentic* (e.g. Archbald & Newman, 1988; Bergen, 1993; Gronlund, 2003; Meyer, 1992; Newman, Brandt & Wiggins, 1998; Wiggins, 1989a, 1989b). The term *best* typically means *valid,* and *authentic* is usually defined as having something to do with the *real world.* This position is difficult to translate into an assessment strategy, however, for two reasons. First, validity is not a characteristic of any assessment; it refers to the interpretation and use of assessment results.

Secondly, there are a variety of definitions of authenticity presented in the research literature and in books and other materials used to train teachers. While most authors speak of authentic in the context of application outside the classroom, some do not and emphasize other aspects of assessments that determine their authenticity. Many advocates emphasize the role of the student in the process or the complexity of the task. Others present criteria that sound suspiciously like general recommendations for valid classroom assessment of any type or, sometimes a bit more specifically, valid *performance-based* assessments of any type. Such recommendations offer little in determining whether any specific teacher-made assessment is authentic and, therefore, produces the benefits presumably associated with authenticity.

Palm (2008). Begins with a basic dictionary definition of authentic as meaning essentially that something is "real, true or what people say it is" (p. 6), he suggests that the term issued in various contexts as being *true* or *real*. Palm concludes that authenticity is defined as assessment that is real in terms of processes and products, assessment conditions or the presented context, and true to life beyond school, curriculum and classroom practice or learning and instruction.

On the contrary in Pakistani educational set up passing the exam is the immediate and the important goal, all instructions in schools and colleges are directed towards that end. Short-cuts are described in learning facts, abbreviated notes and compendiums. It kills the spirit of enquiry and search for truth and gives digested material into the hands of the students which does not constitute a challenge to their mental powers. The teacher is not bothered by the idea of promoting the interest of students to seek knowledge from books, library and laboratory. He himself shows his students the way for an easy approach to the solution by dictating notes.

2.10 Complexity in the assessment of English language proficiency in Pakistani local tests:

English as a second language has a dominant role in the success or failure of any candidate in Pakistani local entry tests. During this research it is observed that usually candidates lament on their failure due to the English language assessment portion in NTS tests. In this research it is observed that the failure ratio is higher in case of the students who gained their elementary education from Urdu medium schools, where English language syllabus is just taught 50%. The sentence parsing and critical writing skills are extremely ignored, the teachers themselves are not sufficiently educated and their own language concepts are either faulty or confused. The lack of adequate pre-service teacher training is frequently observed, such teachers become teachers not by choice but by chance. On the other hand, teachers make complaints of being overloaded with number of classes, due to the increased number of students in English class, teachers cannot be able to pay attention on student's basic understanding of language rules or creative writing skills. Moreover, the teacher himself/ herself cannot get sufficient opportunities to improve his/her concepts or to conduct an extensive research in the language domain. Resultantly, the students from Urdu medium may face extreme dejection while not meeting the higher education criteria.

2.11 Language Test items analysis of local Pakistani Entry test (NTS):

During this study a major issue was highlighted when a teacher dr. Najma from Karachi university put an allegation on NTS, that many of the language test items are frequently plagiarized. This issue reflects NTS negligence as well as the inadequate research in constructing the language test items and lack of standardization which brings incredibility, nonconformity and fakeness to the Pakistani examination systems. There are many candidates, those who make chief complaints about non-standardized test patterns, according to experts non-standardization occurred chiefly due to errors in measurement. This study clarifies that local entry tests in Pakistan (NAT1,NAT2 & GAT) are deficient due to:

2.12 Repetitive usage of past paper samples:

The test patterns are not verified under any International Testing criteria. None of the research is ever conducted to assess the language test items, the candidates cognitive abilities, Assessment difficulty levels, the relationship of the entry test and the candidates academic background. The prescribed curriculum and it's relevance to the local entry tests. The above stated issues are the core issues of this research, and require a satisfactory solution. Before

devising a think tank through this study it is essential that one should have a critical review of the following evaluation criteria which can be help in suggesting few remedies to the Pakistani examiners and local assessment agencies in Pakistan.

2.13 Criterion referenced tests:

Criterion referencing, as applied by professional test agencies, is not directly referring to course objectives or to student learning. Criterion referencing refers directly to test items. A criterion referenced test is one that is prescribed by tight delineations of the structure of particular tasks to be included in the test. Advocates of criterion referenced tests often claim that the performance on such a test is judged in relation to an absolute rather than a relative standard. That is, that scores on criterion referenced tests are measures of achievement in a particular domain and do not depend on relative merit, but are informative in their own right. Criterion referenced scores are in no way absolute scores. They are norm-referenced. The norm-referencing is done prior to the test construction process at the item level, and not at the total test level during a specific application of the test. (Behar 1983, Glass 1978).

Criterion referenced tests contain all of the errors of Mastery tests plus one additional labeling error of great ideological significance. A sub-group of tests in this area, called sometimes Domain referenced tests, have developed a whole theory based on test item characteristics, which is very efficient. Efficient in the sense that students can be tested with less items than in the random sampling model for the same error (an error which, as usual, is never attached to individual scores). This is achieved by using known levels of difficulty of the items (based on random or other specified population estimates).

These kind of tests can highly be effective in case of engineering or science students in Pakistan, e.g. the technical subjects require more attention to be paid on the technical skills, which are required in the advance scientific studies. In this the tests items can be proved more authentic and in relevance to the candidate’s knowledge domain.

2.14 Computing the student's score.

Nothing wrong with this of course. Except the labeling claim that these scores are absolute measures of a "latent trait." What is a latent trait? It is some "hidden characteristic" which some students have more of than others, and which is measured by the test. And those who have more of it are more likely to be able to answer correctly the more difficult items. As all of the items in a Domain referenced test relate to some particular area of learning, such as reading comprehension, or computer skills, or simple calculus, or news paper editing, or social skill, or whatever, then it doesn't really matter what "latent trait" means. The assertion that "it" can be measured absolutely is what constitutes its ideological power. Here is the ultimate rationalization for intellectual and social stratification. Here is the number that describes each person's place on the continuum of ability or skill or whatever for any label that testing agencies wish to attach to the domain of items.

On the surface, of course, it is the specific label that assumes social importance. The claim being made, or at least strongly implied, is that such a test is an absolute measure of reading comprehension, or computer skill etc. But in focusing on the label, we are likely to miss the frightening significance and ideological sleight of hand that produced the "latent trait" as some substantive property or quality permanently attached to the person tested, somehow magically unrelated to the highly subjective, contrived, inter-relational world where a student sits at a desk, reads some questions, and places ticks in computer marked boxes.

Such tests construct current fashionable truths. They are being presented as the latest panacea for testing human ability, or "skills" or "competencies" as they are now called ;they are being presented as the theoretical support for an invasion of competency based assessments in all areas of human measurement (in schools, businesses, bureaucracies ,or where-ever else hierarchies operate). So we should be clear about three things: The first is that constructing a domain referenced test and naming it produces no evidence that the tests measures any sort of trait or ability that can be attached to an individual person (Lord, 1980).The second is that they are not absolute, or error free measures; the scores are related to relative merit, and there is no "standard" performance or score that relates to any minimum or other grade of "competency" that can be theoretically attributed to any score(Glass, 1978).Which takes us to the third point, which is a logical conclusion from the previous two. Domain referenced tests can make little contribution to a field of "competency"
assessment which purports to describe (or more significantly measure) some "standards" of competency in various "skill" areas of human performance.

2.15 Limiting constructs, limiting error:

Let's examine briefly how some of the more general criteria of assessment; labeling, construction, stability, generality, prediction, tend to be limited to what can be controlled by test makers. Labeling is achieved by the simple act of giving a name to the true, or universe, or latent trait, score. Which means, in practice, to the estimated score. The errors implicit in the communication of what that label means, between those who define the course, those who teach it, those who produce the test, those who do it, and those who consume its product, are thus not considered. All of these people will give their various meanings to the label, and make their judgments accordingly. We may be certain that these meanings vary considerably. How much they vary will probably never be known, because it is not in the interests of any institution to uncover yet another source of error. Labeling errors are not currently considered in any estimate of test error. They are immense.

If communication is its effect, then such confusions are, to the student, irrelevant. To the student the meaning of the label is the grade or the mark attached to it. Within the structure that contains the assessment system, the meaning of the label, as distinct from the meaning of the mark, amounts to little more than ideological gossip. At least some students recognize the meaninglessness of the label.

Likewise, construction errors are not estimated; they do enter the theoretical psychometric definitions of validity, but are in practice neither measured nor estimated. The major task of matching objectives to assessment to performance is assumed entirely by the test maker, and most of the errors within this activity are also disregarded, as easily as the errors caused by differing forms of assessment, and use of media other than reading/writing, which don't fit the format of test items on paper, are disregarded. It is assumed that the test is indeed contracted, and the performance required by the student indeed matches, the objectives of the course, or the criterion definitions of the test. Sampling processes that are used, even in professional testing agencies, are at the best primitive, and at the worst nonexistent. This part of test construction is nicely described: as an "art" rather than a science (Nairn ,1980).One thing is certain though; no course has stated as its major, or even minor objective, the ability to answer a pencil and paper test in a given time under stress conditions. And why not? Surely this is the essential behavioral objective.

Stability becomes narrowed to test reliability, more accurately called internal consistency, an internal test measure that cannot take account of variation over time and place and assessors. Theoretically test-retest reliability is one form of reliability, but in practice such estimates are rarely obtained. Generality becomes narrowly construed as related to the extent to which the

test samples the universe of possible test items, or how well the item specifications cover the domain. Generality becomes a function of test items and is called generalizability. Generalizability ignores previous performance in different contexts, forms and media. It ignores all performance other than the purely cognitive response to simulated experience of a multiple choice or written form. It thus ignores all cooperative and all production modes of expression. It reduces human response to the act of recognizing a "best" answer, to conforming adequately to some authority's view of importance, relevance and reality, or to answering someone else's question in a particular way.

And prediction becomes tied to numbers and test scores. In this psychometric world we are no longer concerned with the extent to which actual people are helped to function in differential social situations of great complexity. Prediction does not attempt to describe the relationship between a particular set of learning experiences for some person, and how helpful that is in some future situation for that person. Rather it ranks a group of people on their "success" in the "learning" situation, then ranks them again in some criterion situation. The correlations between the two rank orders represents the predictive value of the test. Not of the course, of the test. And not of its relevance to the quality of their performance, but to its correlation with some person's or group's ranking of their relative performance. And note that even if this correlation is high, which is unusual unless a similar test has been used to measure the criterion, this tells us nothing about whether the relation is in any way causal.

2.16 Validity & Reliability:

The professional theoretical face of assessment discourse asks the question, is the test reliable? More ethically orientated assessors ask the additional question, is the assessment valid? The public wants to know, is it fair? And the more critical of them might add, are people being violated?

Validity:

"Validity," states the first sentence of the APA Standards of educational and psychological testing (American Educational Research Association, 1985), "is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p9). It goes on immediately to explain that: "Test validation is the process of accumulating evidence to support such inferences."Which all sounds very scientific and objective and devoid of bias.

"Invalidity," states the first sentence of the alternative tract, "is the most important consideration in test evaluation. The concept refers to the inappropriateness, meaninglessness, and uselessness of the specific inferences made from test scores. Invalidity or error estimation is the process of accumulating evidence to problematise and ultimately reject such inferences." It should be clear even from this small rewrite that a text that began with the second conceptualization would be a very different text from one that began with the first.

2.17 Positioning:

The main participants in the testing process, we are told, are the test developer, the test user, and the test taker. Also often involved are the test sponsor, the test administrator and the test reviewer. Sometimes, many of these participants may be parts of the same organization, with the notable exception, of course, of the test taker.

The 1985 Standards acknowledge, with fine understatement, that "the interests of the various parties in the testing process are usually, but not always, congruent" (p1). This trivialization of the traumatic effects, dislocations, and exclusions of millions of students based on test and

examination results is quite remarkable. Perhaps it is just another example of the way social positioning can overwhelm interpersonal sensitivity and intellectual honesty.

The concern of the test makers and users is, after all, with hundreds, thousands, or hundreds of thousands of test takers (not to mention their concern with their Board of Directors and shareholders). But their concern is with them, viewed as a group. Their interest is with groups, not individuals; in summaries, not raw data; with simplifying complexities, not with complexifying individuals; with objectifying human subjects, not with subjectifying human events. For the test constructor, sponsor and user there are so many difficult questions; so many criteria to consider; so many factors to consider if the overt and covert claims of the test makers are to be defended. We shall deal with these in due course. Yet to the test taker there is only one question, a normative question which emerges from his or her very construction as an individual. Have I passed or have I failed? Am I satisfactory or unsatisfactory? Am I normal or a nut case?

Additionally and ironically, it is precisely because they see the testing event from this Individualized perspective, rather than from a group perspective, that they do not ask the more crucial, the more fundamental question: How much error, ambiguity, uncertainty, does this attribution contain? Or is it their powerlessness, and unheard voice, that makes these questions at the best unspeakable, at the worst unthinkable?

2.18 Sources of evidence:
The 1985 International assessment Guidelines described an ideal validation as including several types of evidence. . . Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable validity (p9).

This is hardly reassuring for the test taker. The tautology and redundancy in the phrase "questionable validity" is remarkably inept; validity is proposed as the characteristic of the evidence used to support the construct "validity," and the essence of the concept is surely its very questionability. Far more damning, however, is the clear implication that evidence that does not cogently support the assertions of the test users should not be presented. Putting it another way, validity is a concept based on advocacy, is a rationalizing tool for a methodological decision already made, and is an ideological support rather than a scientific enterprise. Is this an over-statement? Here is the first sentence of the next paragraph of the 1985 Standards: "Resources should be invested in obtaining a combination of evidence that optimally reflects the value of a test for an intended purpose" (p9). The word "optimally" says it all. So, validity is clearly an advocacy construct, based on the assumption that any assessment data is innocent until proved guilty. The discourse about validity presents the case for the defense.

Yet here we also see, in the very heartland of post-positivist empiricism, the embryo of a discursive construct; an appeal, not to numbers, but to discourse. Over the next ten years Cronbach (1988) and Messick (1989a, 1989b,1994),doyens of psychometrics, in their born-again personas will enlarge the idea of construct validity to a point where Cherryholmes (1988) will nail it as fully discursive, and thus "linguistically, politically, economically, socially, culturally and professionally relative"(p450).

Even so, the advocacy position remains essentially unchanged. Messick(1989b)
asserts that :

“To validate an interpretive inference is to ascertain the extent to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported. This represents the fundamental principle that both convergent and discriminate evidence are required in test validation (p1). But note the implication of "are less well supported" and its relationship to advocacy. And later in the same article, when he gets specific about invalidity implications of adverse social consequences, he says: If the adverse social consequences are empirically traceable to sources of test invalidity, . . . then the validity of test use is jeopardized. . . If the social consequences cannot be so traced - or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible - then the validity of the test use is not overturned.”

Note the use of the words "jeopardized," "less plausible," and "not overturned." Given the probabilistic nature of all social research, the chances of any test being declared invalid on the basis of these criteria, from this perspective, are remote. Ultimately, Messick is eminently logical. For if "validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(Messick, 1989a, p13), then even infinitesimal support, being support, makes the test valid, and nothing has really changed since Guilford's (1946) claim that "in a very real sense, a test is valid for anything with which it correlates " (p429). And as error will ensure that no tests correlate zero with anything, it follows that all tests are valid.

2.19 Reliability:

Even though validity has taken on a post-modernist hue of recent times, reliability has, until recently, remained untouched as a "foundational" cornerstone of educational measurement. Reliability was seen as the lower limit of validity. An assessment could not be more valid than it was reliable. The assessment industry, whether local, corporate, government, has embraced the reliability concept both ideologically and empirically. In contrast to validity, estimates of reliability are often obtained and circulated. There are two reasons for this: the reliability of the test can be measured using only data from the test scores; and often relatively high values (correlations of 0.7 - 0.9) can be obtained, if for no other reason that they are so constructed to ensure that such high internal consistency occurs.

Politically such reliability data can be used to "prove" the quality of the test, and maintain the illusion that reliability refers to "the degree to which test scores are free from errors of measurement," which is how they are described in the first sentence about Reliability in the 1985 Standards. In fact, the Standards emphatically insist that: For each total score, sub score, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test (p20).

Note that it is never suggested that the standard errors of measurement information should be available to test takers. There is a later chapter in the 1985 Standards entitled "Protecting the rights of Test Takers." Again there is not the vaguest suggestion here that such information should be made available to them. However, even reliability is now under threat. Is there nothing sacred? Moss (1994), has cogently argued that there can be validity without reliability. She points out that: Reliability, as it is typically defined and operationalised . .privileges standardized forms of assessment. By considering hermeneutic alternatives for serving the important epistemological and ethical purposes that reliability serves, we expand the range of viable high-stakes assessment practices to include those that honor the purposes that students bring to their work and the contextualized judgments of teachers (p5).

Such idiosyncratic behaviors and judgments tend towards a diversity that reliability abhors. There are two issues here. The first relates to the relationship between reliability and validity perceived from the standpoint of the assessors; the second deals with the concept of reliability, that is consistency, of performance as actually produced by the persons being assessed. The two issues are related in that they both relate to responses to persons involved in an event designed to describe what a person can do by asking them to do something else, and then making inferences about what they might do in another time and place and context.

Let's first look at this expectation of high reliability, and the theorizing that precedes it. The argument is essentially this - if one test or examination is reliable then another similar test or examination will give the same verdict, however that verdict is communicated - as marks, grades, pass-fail, selected, or whatever. It is logical to assume, therefore, that one half of the test would give the same verdict as the other half, because all of the bits of the test contribute to the final score and hence the final verdict; putting it another way, we are dealing with some linear dimension here, some unitary idea or construct; all of the questions measure it with considerable error, but the more interconnected questions we ask, and the more inter-correlated answers we get, the more the error is reduced, and the more the measurement is refined to approach the true measure of it. Of one thing we are sure. The "it" is out there, waiting to be measured And "it" has a true value, that we can approach but never completely determine. This simplistic positivism is at the epistemological and ontological heart of educational measurement. Teachers and public examination boards do not believe that this is what they are doing, even though the latter have no hesitation is using measurement theory to manipulate their results and rationalize their processes. They do not necessarily believe there is some unilateral trait or ability or skill that underlies the total score or grade.

Indeed, as Willmott and Nuttall (1975) point out: In the field of 16+ examining it is quite possible that any increase in reliability would be to the detriment of validity. This is easily seen to be the case, since by refining questions and components so that they correlate highly is to learn more and more about less and less: the trait being measured is defined even more narrowly as reliability (in the sense of internal consistency) is increased. In such a situation, the validity of the examination concerned is bound to decrease owing to the narrowness of the field covered. A glance at any subject syllabus published by a CSE or GCE board shows clearly that the comprehension of a very wide variety of content is required of candidates and, in many cases, the educational objectives required of candidates in following the course are equally varied.

It is a pity that these authors do not take this argument to its logical conclusion:
that there is no single trait to be measured, that there is no linear concept to be categorized, and that there is no necessary correlation - indeed there may be some negative correlations, between the relative performances of candidates on various objectives. But this conclusion would lead inevitably to the final one, that there can be no meaningful rank order of students, because the rank order can give no meaningful information about the performance of individual students in relation to any particular objective. Incidentally, correlations across different subjects are often also of the order of0.8. That is the correlation between two tests of different subjects is about as high as the reliability of any one test. (quoted by Nuttall & Willmott, 1975).

Perhaps there is a linear trait after all, but unrelated to the apparent construct being measured. What might this construct be? Traditionalists would be in no doubt that it was a general ability that they would label intelligence. Yet we know that the correlations between examination scores and other sorts of measures (e.g. job performance) are very low, of the

order of 0.3. So a more direct and sustainable interpretation is that "it" is the ability to perform in the events constructed around examinations. Examinations measure examination ability! Now this argument, if we take it a little further, leads to a very strange conclusion. Let's go back to the lines of the Willmott and Nuttall (1975) quote: "it is quite possible that any increase in the reliability would be to the detriment of validity".

They show why this is so in the measurement of any multi-dimensional area, and Moss (1994) indicates why it is so for "hermeneutical alternatives." But increase in reliability from what point? From0.8, or from 0.5 ? Or from zero? Is there an argument to be made that all reliability negates validity. This would lead us to the apparently absurd conclusion that the greater the reliability the lower the validity, and the ultimately maximum validity is to be obtained from zero reliability. In terms of measurement, this would mean, of course, that human "constructs" were essentially immeasurable. We can talk about them, but we can't measure them.

Contextual errors are certainly increased by confining assessment to pencil and paper situations and producing a very singular and artificial environment in which the assessment occurs, to the extent of standardizing format and time available to complete the tasks. Again reliability is obtained at the expense of validity, which implies generalizing to other contexts. Construct errors are likewise increased through the limitations of content, form, process and media that is determined and narrowed through the testing or examination procedures. Again the capacity to generalize, and thus the validity, is diminished by the psychometric strictures required for high reliability.

In similar vein, errors attributable to frame of reference shifts, to labeling and attachment confusions, to prediction inaccuracies, or to logical type confusions, are largely indifferent to reliability. And whilst consequential errors, the negative effects of testing, have certainly been exacerbated by the quest for higher reliability, it is the quest rather than the empirical value that is involved. Instrumental errors of course are reduced as reliability increases; indeed, reliability may be defined as the inverse of instrument error. So in this one area it is clear that increases in validity are dependent on increases on reliability. Yet if, as we have shown, the effect elsewhere is that such increase in reliability either decreases validity or has an indeterminate effect on it, then the general proposition holds, and we may say that in the empirical world, the procedures used to increase reliability result in a decrease in validity.

Messick (1989a) has broadened the concept of validity to refer to "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(p13), and this includes the way values "influence in more subtle and insidious ways the meanings and implications attributed to test scores"(p59), so that "test validation embraces all of the experimental, statistical and philosophical means by which hypotheses and scientific theories are evaluated" (p14).

Messick's position seems to be generally accepted. The sources of potential error actually referred to do cover the range and depth of epistemological, ontological, and value sources referred to in this thesis. Yet even with this multiplicity of error, this proliferation of possibility of miscategorization. Messick (1989) insists that validity is a unitary concept, a singular "degree of support": The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force behind this integration is the trustworthiness of empirically grounded score interpretation, that is, construct validity.

In other words, validity is a statement of faith in testing, a statement of justification by an" expert" that the whole assessment event is legitimate, is valid. Even though, in practice, for real tests, the considerations and scientific inquiries that Messick advocates are rarely carried out. Let's look at this in more detail; first it is apparent that appropriateness, meaningfulness, the usefulness are sometimes quite separable. Appropriateness applies very much to particular values. In my value system, any test which violates individual students is inappropriate. Yet it might be quite meaningful in that some inferences made from it can be understood and acted on by teachers and administrators, and it may be useful in that predictions made from it help selection processes. In another case a test of inverted neuroticism may be quite useful in predicting successful medical students, but may be considered inappropriate for that application. It's meaningfulness may be moot. Ultimately, of course, the very meanings of appropriate, meaningful and useful are deferred; they are partial synonyms for valid, the word they supposedly elucidate.

It becomes clear that the "unifying force" then is not created by the congruencies among appropriateness, meaningfulness and usefulness, but rather by the "trustworthiness" of the "interpretation." In other words, by the power that resides in the status of the "expert" who controls the discourse in which the judgment is embedded. And because the discourse of validity is in essence about all the ways in which the measurement cannot do all the things it claims to do, and explicitly about some of the ways it might be done better, an advocacy judgment would concentrate on some way or ways in which the test was better than it might have been had such improvements not been made.

According to Messick, this is the unifying force that asserts, and thus proves,
validity. Messick's (1989a) definitive paper in the third edition of Educational measurement indicates that he makes reference to over fifty sources of potential invalidity; for indeed, how can he describe how a test may be valid without focusing on all of the ways in which it might not be valid.

Finally, the very existence of validity is established, validity is indeed made manifest, through the denseness of the arguments used to refute such existence, together with the reassurance that the battle continues, and some gains have been made. Let me be specific: The definition of the construct of validity does not exclude the notion of invalidity. However, the discourse on validity, constructed as it is from the position of advocacy, excludes the notion of invalidity as an issue. More than this, the discourse itself becomes the arbiter of the proof of validity claims, independently of empirical data, which becomes irrelevant within the density and complexity of the discourse; as a result, empirical data to justify validity claims is rarely collected, and when it is it is inevitably construed as supporting the claim. Evidence rejecting the validity claim is never collected because such positioning is absent from the discourse.

Madaus (1986) puts it nicely: present methods of gathering content validity evidence are inadequate; they are designed in such a way as to almost guarantee a positive outcome. Alternative methods designed to disconfirm or test counter hypotheses about the issues are, in my experience, never employed (p12).

Practically, the psychometric scam is accomplished by focusing on the test score, and ignoring its dark side, the standard error of estimate; specifically, by implying that the estimated score is the true score, that the intention is the empirical fact, that talking about problems of validity magically increases validity, and that increasing validity makes a test valid.

2.20 Validity and the predominant paradigm:

When advocacy is positioned, aligned to the predominant paradigm, then advocacy is interpreted as truth. Truth not as the production of true utterances, but in Foucault's (1982) sense of "the establishment of domains in which the practice of true and false can be made at once ordered and pertinent"(p8). From the 1980s, when the prevailing societal metaphor is the discourse that surrounds economic rationalism, and in particular those myths connected with people competencies, the metaphor is rapidly post-positivist, and validity definitions (advocacies) based on those assumptions will be seen as self-evidently true. As Cherryholmes (1988) puts it from his post-modern perspective: "boundaries limiting construct-validity discourse have yet to be justified. They are policednonetheless "(p154).In contradistinction, advocacies for more post-modern descriptions (eg validity characteristics for qualitative research) are clearly not aligned to the prevailing world-view, and so will be interpreted as justifications. They advocate from a loser's position, so at the best their views are accepted as tentative, at the worst as unproven and hence unacceptable assumptions. This is inevitable because no abstraction can be proven to be correct, so acceptance is always a function of value, rather than of rational proof; and moral value is usually construed as stabilization of the status-quo, as confirmation of the predominant paradigm. Shepherd (1991) gives an example: "measurement specialists asserted that performance assessments are less reliable and less valid than traditional tests and that they are potentially biased because they rely on fewer tasks." But then she adds: "Why are existing tests presumed to have the high ground in this dispute? What claim do traditional tests have to validity?" . This is not to deny the acceptance of such advocacy in localized communities e.g.(some faculties of some Universities) where a paradigm shift has already occurred.

2.21 Qualitative assessment and qualitative research:

Validity criteria in qualitative assessment has lagged behind validity in quantitative research. However, the two fields are closely aligned. In fact Messick (1989) regards then as virtually synonymous in that test validation in essence is scientific inquiry into score meaning -nothing more, but also nothing less. All of the existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminate arguments to construct interpretation of test scores" .

This study pinpointed various intricate issues in the assessment procedures in Pakistan, it justifies the logical reasons of errors in measurement which become an ultimate cause of non-standardization of the test items, the study has focused manifold challenges to local testing mechanism, which really require improvement, this could only be possible if sound research and massive surveys will be conducted to bring reformation in the Pakistani examination systems.

In particular, this study indicated how the notion of advocacy built into the very definition of validity overwhelms scientific detachment, and effectively silences the logical inferences that derive from the voices of confusion and error that are the very basis of validity discourse. The emphasis on reliability of assessment instruments is also shown to be a misplaced source of credibility for assessment, because measures to increase reliability are shown to decrease validity. Now the coin can be flipped. The underside of validity can be examined.

The extensiveness was necessary, as many of the misconceptions and fudges and contradictions that characterize the field of educational assessment have been caused by a myopia regarding

knowledge outside the arbitrary boundaries within which the field encloses itself. Within the field of educational measurement the critical studies which most Overlap are: in the United Kingdom, Hartog & Rhodes (1936), Cox (1965); in the United States, Hoffman (1964), Nairn (1980), Airasian (1979), and Glass (1978); in Australia, Rechter & Wilson (1968).The Hartog & Rhodes study clearly showed the enormous instability of the measurement of standards in Public Examinations in England.

The sneakiness of some of the research techniques in no way detracts from the dramatic incisiveness of the data. Cox did a similar job and ended up with a similar horror story on measurements of University grades. Hoffman directed his critical attention to the detail of multiple choice testing. Nairn's critique of the work of Educational Testing Service, and in particular the part it plays in College Entrance, is devastating in its implications. Airasian's book is a comprehensive critique of competency testing. Glass attacks the measurement of standards at its most vulnerable point; there are no standards, or at least none that psychometrics can produce. And Rechter & Wilson's study indicates the confusion about how to reduce error that accompanies public examining in Australia.

On the other hand, most of the literature on reliability and validity is pertinent to this study, because, when its discourse is repositioned from examiner to examined, it provides more than enough invalidity information to self destruct. Most studies of error in the measurement of standards are however much more specific in their focus than is mine. Their minimal effect on practice has perhaps partially been due to the fact that their critiques were in terms of their own discipline of educational measurement; a discipline that owes its very existence to the claim to accurate judgments. In terms of general style and scope this study is perhaps closer to the work of Persig (1975; 1991), who delved, articulately if deviously, much more deeply into the notion of quality.

Within the field of power relations and the construction of the individual the studies most similar are those published in Foucault and Education (Ball,1990), in particular those that take off from Foucault's placement of the examination as a central apparatus of power/knowledge. This study is significant in that it brings these two diverse fields of educational assessment, and the power relations that pervade education, into much closer contact, to expose their interrelations, and allow the critique to cross fertilize.

This review of literature emphasizes that language tests should mediate ideologies and practices in more closer manner, it should be more democratic and negotiable rather than tests as powerful mechanism capable of imposing draconic educational policies without having in empirical base. According to Davies (1997), language tests can become a useful tool for negotiating between language ideologies and language practices.

Chapter 3:

3.0Research Methodology:
This research is descriptive/survey based analysis of the current scenario of summative and formative assessment practices prevalent in Pakistani education system: Following research procedures were adopted to probe into the non standardized examination practices existing in Pakistan.

3.1Population Sample:
Two public sector and two private sector universities were selected to gather the data through the survey; the English language teachers and students constituted the population of this study. In this survey one hundred students and forty English language teachers participated from the four renowned universities of the city of Lahore, district Punjab; Two universities were private universities and the other two were public sector universities.

The research survey was gathered from the following universities:
University-wise no. of teachers and students is given below:

Universities	Teachers	Students
Public university 1	10	25
Public university 2	10	25
Private university 1	10	25
Private university 2	10	25

The data was collected from these universities, two questionnaires consisting of 24 question items were distributed amongst the students and English language teachers of these universities.

3.2 Research instruments:
After going through exhaustive literature review, experts' feed back, policy documents & reports ,the researcher prepared two questionnaires. A questionnaire having 45 items for university English language teachers both at public & private sector. A questionnaire having 24 items for university students both at public & private sector. These questionnaires were based on the five point Likert Scale 5.0.

3.3 Pilot testing:
For examining the understandability of the question items in this survey. 2 universities, both public & private sector universities were selected initially, 10 English language teachers and 30 students were given these questionnaires for trialing the credibility of the question items of this survey. This piloting was conducted by the researcher for the pre-screening of the question items in order to make necessary amendments before undergoing the main survey. The respondents were asked to be critical about each questionnaire's language & construction. They were given freedom to make amendments according to their ability to comprehend the questions.

3.4 Data Analysis: The Software SPSS 17.0 was used to make the data analysis more accurate. The results were gathered through percentile and correlation Methods. This research is based on four variables upon which the subsets of questions were formulated; each set of questions has given a specific code in order to find out the results through SPSS 17.0. In order to show significant relation of the research objectives the Pearson Co-relation was applied on the results gathered through the research.

3.5 Research Variables: The research had four objectives upon which the two questionnaires were based. One objective was to probe the English language teacher's point of view regarding the examination practices both at the summative and formative levels, the examination strategies and assessment mechanisms. There were thirty five question items including the 10 interview questions. The second questionnaire was designed to highlight the students' points of view regarding examination policies, academic bench marks, reformations & also their concerns about the difficulty levels related to both the summative and formative assessments.

The following data sets were carefully designed based on several question clusters, which were grouped and coded for SPSS 17.0 analysis. The data sets are attached with each research objective categorically:

Data set 1: Research objective 1:
1.To highlight the non-standardized evaluation mechanism in Pakistan. (SPSS Data Code: NEP, Non-standardized Evaluation Practices)

Question cluster under objective 1. Code: NEP.

1).Do you think that NTS Tests have really contributed in improving your English language skills?

2.Do you think that the NTS Tests are merely a formality & do not focus on your university/college English language needs?

10).Do you think that the English language test pattern of NTS is not focusing on your creative writing ability?

11).Do you think that the English language test pattern of NTS is not sufficient to prepare you for any International language proficiency tests like IELTS, TOEFL,GMAT or GRE?

24). Do you think that the local English language test would be beneficial for the Pakistani students in future?

Data set 2: Research objective 2:
2. To assess examination standards both at the formative and summative levels in Pakistan. (SPSS Data Code:(FSM: Formative/Summative Assessments)

Question cluster under objective 2. Code: FSM (Formative/Summative Assessments).

12). Do you think that the English language courses in your university curriculum really fulfill your academic needs?

15). Are you satisfied with the grade you get in your English language assessments in class?

16). If you do not get good grades in the English language tests, is it because of your poor previous English language background?

17). Do you think that your not getting good grades is because of the boring curriculum scheme?

18). Do you think that your not getting good grades is because you lack the skills required for attempting questions?

19). Do you think that your not getting good grades, is because you find the multiple choice questions difficult?

20). Do you think that your not getting good grades is due to infrequent practice?

21). Do you think that your not getting good grades is due to ineffective reading & writing skills?

Data set 3: Research objective 3:
3.To identify the inefficacy of local testing bodies, which is a cause of de-motivation amongst Pakistani students. (SPSS Data Code:(INT).
Question cluster under objective 3. Code: INT(Inefficacy of National Testing Systems).

3). Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?

4). (If you disagree with question no 3) do you think that NTS is a reliable & transparent assessment system?

5). Are you satisfied with the assessment score awarded to you by NTS?

13). Do you think that due to NTS entry tests, it is difficult to obtain admission in universities?

14). Do you think that every university must conduct it's own English language entry test?

22). Do you think that the NTS entry test language portion is inadequate because it does not assess listening and speaking skills?

Data set 4: Research objective 4:

To identify the of lack of unified Marking assessment criteria of NTS. (SPSS Data Code: (NAC).

6).You are not satisfied because you think transcript was not properly checked?

7).You are not satisfied because you think English language assessment was too difficult?

8).You are not satisfied because you think that your numerical ability is stronger then your language ability?

9).Do you think that English language multiple choice questions are sufficient to assess your English language proficiency?

23). Do you think that there should be a separate local English language test, as university entry criteria?

The above stated question clusters were carefully designed to conduct a survey while utilizing 5 point Likert Scale; the responses were gathered according to the following sequence

Strongly Agree	5
Agree	4
Neutral	3
Disagree	2
Strongly Disagree	1

3.10 Likert Scale : The Likert Scale was adopted to gather the respondents' preferences or degree of disagreement with a statement or set of statements. In this way, the researcher came across loop holes in the functioning of the local examining bodies, and also highlighted the necessary steps which could be taken on the behalf of the education ministry. Therefore, Likert scale proved to be an appropriate method to access the opinions of the Pakistani students & teachers. It is a psychometric response scale through which the respondents are asked to indicate their level of agreement with a given statement by the way of an ordinal scale.

3.11 Data Interpretation: The survey which was conducted to gather the opinions about the efficacy of testing and evaluation systems in Pakistan. The following tables exhibit the results of the survey which was conducted in 4 distinctive universities of Lahore.
The statistics which are illustrated in table:3.11.0 to table:3.17.4 reveal whether non-standardization is prevalent in local examination systems or not? The following results exhibit the respondents' feedback about the non-credibility of evaluation mechanisms in Pakistan.

3.11.0 Objective 1. To highlight the non-standardized evaluation practices in Pakistan.

Table 3.11.0: Do you think that NTS Tests have contributed in improving your English language skills?

NEP1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	22	22.0	22.0	22.0
	2	23	23.0	23.0	45.0
	3	24	24.0	24.0	69.0
	4	24	24.0	24.0	93.0
	5	7	7.0	7.0	100.0
	Total	100	100.0	100.0	

According to table (3.11.0), 93% respondents are in favour that if the National testing system in Pakistan gets more scrutinized than it can become a vital source of students improvement in English language skills in Pakistan.

Table 3.11.1: Do you think that NTS Tests are merely a formality & do not focus your university/college English language needs?

NEP2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	10.0	10.0	10.0
	2	27	27.0	27.0	37.0
	3	23	23.0	23.0	60.0
	4	28	28.0	28.0	88.0
	5	12	12.0	12.0	100.0
	Total	100	100.0	100.0	

According to table 3.11.1 88% respondents are of this opinion that National testing system in Pakistan is merely a formality and it has nothing to do with their basic academic requirements in university or college because the formative level of English language assessments is comparatively higher than the local level assessments.

Table 3.11. 2: Do you think that English language test pattern of NTS is not sufficient to improve your creative writing ability?

NEP3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	15.0	15.0	15.0
	2	19	19.0	19.0	34.0
	3	15	15.0	15.0	49.0
	4	38	38.0	38.0	87.0
	5	13	13.0	13.0	100.0
	Total	100	100.0	100.0	

According to table: (3.11.2) 87% respondents were of this opinion that the NTS test pattern in case of English language assessments has no contribution in improving their creative writing skills because the examination format is more objective.

Table 3.11.3: Do you think that English language test pattern of NTS is not sufficient to prepare you for any International language proficiency tests likewise: IELTS, TOEFL,GMAT or GRE?

NEP4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	9.0	9.0	9.0
	2	15	15.0	15.0	24.0
	3	19	19.0	19.0	43.0
	4	31	31.0	31.0	74.0
	5	26	26.0	26.0	100.0
	Total	100	100.0	100.0	

According to table (3.11.3), 74% respondents showed their dissatisfaction with the standard of the National Testing System in comparison to the International examination systems. Therefore, the local testing bodies are not considered as benchmarks by the Pakistani students.

Table 3.11.4: Do you think that local English language test would be beneficial for Pakistani students in future?

NEP5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	12	12.0	12.0	12.0
	2	8	8.0	8.0	20.0
	3	28	28.0	28.0	48.0
	4	25	25.0	25.0	73.0
	5	27	27.0	27.0	100.0
	Total	100	100.0	100.0	

According to table: (3.11.4),73% respondents desired improvement in the NTS test patterns; according to the respondents, it can become a sound benchmark in future, if the test patterns are designed according to the International standards.

3.12 Objective 2.To assess examination standards both at formative and summative levels in Pakistan. (SPSS Data Code:(FSM: Formative/Summative Assessments)

The below tables (3.12.0) to (3.12.7) illustrate the results which are gathered in order to assess examination standards both at formative and summative levels in Pakistan. The below given tables assess the students feedback about the formative and summative assessment practices.

Table:3.12.0. Do you think that English language courses in your university curriculum really fulfill your academic needs?

FSM1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	12	12.0	12.0	12.0
	2	26	26.0	26.0	38.0
	3	17	17.0	17.0	55.0
	4	32	32.0	32.0	87.0
	5	13	13.0	13.0	100.0
	Total	100	100.0	100.0	

According to Table: (3.12.1) 87% respondents showed their satisfaction about the standards of English language syllabi prescribed in their universities; they believed that the syllabus is meeting the academic requirements of the respondents, and through the prescribed scheme of study they can easily compete at the International level.

Table:3.12.2 Are you satisfied with the grade you get in your English language assessments in class?

FSM2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	7.0	7.0	7.0
	2	12	12.0	12.0	19.0
	3	17	17.0	17.0	36.0
	4	45	45.0	45.0	81.0
	5	19	19.0	19.0	100.0
	Total	100	100.0	100.0	

Table (3.12.2) illustrates that 81% respondents are satisfied by the grades they get in the subject of English language at university level. They believed that they really need to work hard in order to have satisfactory grades; furthermore, they responded that the English language course in their universities is really effective in improving their background knowledge.

Table:3.12.3 You do not get good grades in English language tests, it is because of your poor previous English language background?

FSM3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	15.0	15.0	15.0
	2	24	24.0	24.0	39.0
	3	19	19.0	19.0	58.0
	4	27	27.0	27.0	85.0
	5	15	15.0	15.0	100.0
	Total	100	100.0	100.0	

According to table (3.12.3) 85% respondents agreed that they have poor academic background due to insufficient instructional base. They also showed their dissatisfaction from the so called English medium schools and also the dual standards likewise: Urdu & English Mediums.

Table:3.12.4 You do not get good grades because you think curriculum scheme is boring?

FSM4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	17	17.0	17.0	17.0
	2	26	26.0	26.0	43.0
	3	26	26.0	26.0	69.0
	4	23	23.0	23.0	92.0
	5	8	8.0	8.0	100.0
	Total	100	100.0	100.0	

According to Table (3.12.4) 92% respondents from Urdu medium background were of the opinion that the prescribed curriculum has no innovation, the syllabus needs frequent revisions.

Table:3.12.5) You do not get good grades, it is because you think that you lack skills required for questions attempting?

FSM5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	10.0	10.0	10.0
	2	18	18.0	18.0	28.0
	3	20	20.0	20.0	48.0
	4	36	36.0	36.0	84.0
	5	16	16.0	16.0	100.0
	Total	100	100.0	100.0	

According to table (3.12.5) 84% respondents agreed that they lack question attempting skills, therefore, they have poor grades in English Language according to the university/college assessments.

Table:3.12.6 You do not get good grades, it is because you think that multiple choice questions are difficult?

FSM6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	19	19.0	19.0	19.0
	2	34	34.0	34.0	53.0
	3	29	29.0	29.0	82.0
	4	11	11.0	11.0	93.0
	5	7	7.0	7.0	100.0
	Total	100	100.0	100.0	

According to the above table(3.12.6) 82% agreed that they really have had little exposure to the multiple choice questions attempting techniques, which is considered an important part of English language objective tests both in the entry level tests and in-house examinations.

Table:3.12.7 You do not get good grades, it is because you are not provided with frequent practice?

FSM7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	10	10.0	10.0	10.0
	2	26	26.0	26.0	36.0
	3	28	28.0	28.0	64.0
	4	28	28.0	28.0	92.0
	5	8	8.0	8.0	100.0
	Total	100	100.0	100.0	

According to table (3.12.7) the respondents admitted their own negligence for not preparing well; they considered lack of appropriate and frequent practice as the major cause of their low grades in English language tests.

Table:3.12.8 You do not get good grades, it is because you think you lack effective reading & writing skills?

FSM8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	15.0	15.0	15.0
	2	21	21.0	21.0	36.0
	3	24	24.0	24.0	60.0
	4	29	29.0	29.0	89.0
	5	11	11.0	11.0	100.0
	Total	100	100.0	100.0	

Table:3.12.8 highlights that 89% respondents admitted that the cause of their poor proficiency is their lack of good reading and writing skills; they admitted their reluctance in improving their language proficiency skills.

3.13.0 Objective:3. To identify the inefficacy of local testing bodies, which is a cause of de-motivation amongst Pakistani students. (SPSS Data Code:(INT).

Table:3.13.0 Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?

INT1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	18	18.0	18.0	18.0
	2	41	41.0	41.0	59.0
	3	13	13.0	13.0	72.0
	4	22	22.0	22.0	94.0
	5	6	6.0	6.0	100.0
	Total	100	100.0	100.0	

According to Table:3.3.0 94% respondents strongly agreed that the entry test autonomy must not be given to the universities/colleges because they might set their own biased criteria, which may lead to unfairness in assessment. The respondents strongly preferred improvements in the NTS as an objective examination body.

Table:3.13.1 If you disagree then do you think that NTS is a reliable & transparent assessment system?

INT2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	9.0	9.0	9.0
	2	23	23.0	23.0	32.0
	3	23	23.0	23.0	55.0
	4	31	31.0	31.0	86.0
	5	14	14.0	14.0	100.0
	Total	100	100.0	100.0	

Table:3.13.1 clearly shows that 86% respondents are in favor of an unanimous and transparent examination body which should be highly mechanized and must have a central authority as far as admissions are concerned.

Table:3.13.2 Are you satisfied with the assessment score awarded to you by NTS?

INT3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	7.0	7.0	7.0
	2	19	19.0	19.0	26.0
	3	19	19.0	19.0	45.0
	4	43	43.0	43.0	88.0
	5	12	12.0	12.0	100.0
	Total	100	100.0	100.0	

Table: 3.13.2 reveals that respondents themselves have poor background in English language skills; therefore 88% respondents agreed that their NTS scores were justified to a great extent.

Table:3.13.3 Do you think that due to NTS entry tests, it is difficult to obtain admission in universities?

INT4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	15.0	15.0	15.0
	2	32	32.0	32.0	47.0
	3	32	32.0	32.0	79.0
	4	14	14.0	14.0	93.0
	5	7	7.0	7.0	100.0
	Total	100	100.0	100.0	

Table:3.13.3 reveals that 79% respondents consider NTS as neither difficult nor easy; the respondents were of this opinion that all parts of the question paper are interrelated therefore, Numerical attempt is equally required to be scored well as English language portion in NTS.

Table:3.13.4 Do you think that every university must conduct it's own English language entry test?

INT5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	16	16.0	16.0	16.0
	2	14	14.0	14.0	30.0
	3	22	22.0	22.0	52.0
	4	33	33.0	33.0	85.0
	5	15	15.0	15.0	100.0
	Total	100	100.0	100.0	

According to Table: 3.13.4, 85% respondents are of this opinion that the universities/colleges must conduct an English language entry test which is beneficial for the students and also the institutional authorities must have an idea about the strengths and weaknesses of the pupil they are inducting.

Table:3.13.5 Do you think that NTS entry test language portion is inadequate, because it does not assess listening and speaking skills?

INT6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	5.0	5.0	5.0
	2	15	15.0	15.0	20.0
	3	22	22.0	22.0	42.0
	4	34	34.0	34.0	76.0
	5	24	24.0	24.0	100.0
	Total	100	100.0	100.0	

In Table:3.13.5 76% respondents had reported that listening & speaking skills are not assessed in any local entry test in Pakistan; therefore the pupils remain ignorant about their linguistic deficiencies in this respect and they also suffer in future due to these linguistic in competencies.

Table: 3.13.6 You are not satisfied because you think transcript was not properly checked?

NAC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	16	16.0	16.0	16.0
	2	36	36.0	36.0	52.0
	3	29	29.0	29.0	81.0
	4	15	15.0	15.0	96.0
	5	4	4.0	4.0	100.0
	Total	100	100.0	100.0	

Table: 3.13.6 reveals the neutrality of the respondents' point of view about transcript checking. The results also reveal partiality of the respondents' feedback.

Table:3.13.7 You are not satisfied because you think English language assessment was too difficult?

NAC2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	12	12.0	12.0	12.0
	2	37	37.0	37.0	49.0
	3	27	27.0	27.0	76.0
	4	20	20.0	20.0	96.0
	5	4	4.0	4.0	100.0
	Total	100	100.0	100.0	

Table:3.13.7exhibits that 76% respondents were of this opinion that English language assessments are neither difficult nor easy. Therefore the respondents' points of view vary greatly.

Table:3.13.8 You are not satisfied because you think that your numerical ability is stronger than your language ability?

NAC3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	9	9.0	9.0	9.0
	2	25	25.0	25.0	34.0
	3	32	32.0	32.0	66.0
	4	23	23.0	23.0	89.0
	5	11	11.0	11.0	100.0
	Total	100	100.0	100.0	

According to Table:3.13.8, 89% respondents reported that their numerical ability is far stronger than their linguistic ability. Furthermore, they do not consider English as a "lingua franca"; therefore, they are not interested to learn English as a second language. The respondents reported that they feel improving their English language skills a burden, whereas numerical questions interest them and they can score well in numerical portions.

Table:3.13.9 Do you think that English language Multiple choice questions are sufficient to assess your English language proficiency?

NAC4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	15	15.0	15.0	15.0
	2	27	27.0	27.0	42.0
	3	13	13.0	13.0	55.0
	4	33	33.0	33.0	88.0
	5	12	12.0	12.0	100.0
	Total	100	100.0	100.0	

According to Table:3.13.9, 88% respondents believed that the multiple choice questions are not considered sufficient to assess the candidates in the entry test because students take an edge of over attempting the options, moreover, MCQ'S cannot really assess the critical & creative writing skills of the candidates.

Table:3.13.10 Do you think that there should be a separate local English language test, as university entry criteria?

NAC5

	Frequency	Percent	Valid Percent	Cumulative Percent
1	8	8.0	8.0	8.0
2	28	28.0	28.0	36.0
3	15	15.0	15.0	51.0
4	35	35.0	35.0	86.0
5	14	14.0	14.0	100.0
Total	100	100.0	100.0	

According to Table:3.13.10, 86% respondents were of this opinion that their skills in English as a second language must be assessed locally as these are assessed Internationally in TOEFL & IELTS. In this way, the local English language tests would help Pakistani students to compete internationally.

3.14.0 Data Analysis of the English language teachers:

Objective1.To highlight the non-standardized evaluation mechanism in Pakistan.(SPSS Data Code:(NEM)

Question cluster under objective 1. Code: NEM

1). Do you think that NTS Tests have contributed in improving your students English language skills?

2). Do you think that NTS Tests are merely a formality & do not focus students university/college English language needs?

10). Do you think that English language test pattern of NTS is not focusing student's creative writing ability?

11). Do you think that English language test pattern of NTS is not sufficient to prepare them for any International language proficiency tests likewise: IELTS, TOEFL,GMAT or GRE?

24). Do you think that local English language test would be beneficial for Pakistani students in future?

(Table: 3.14.0) Do you think that NTS Tests have contributed in improving your students English language skill.

NEM1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	4	17.4	17.4	17.4
	2	8	34.8	34.8	52.2
	3	6	26.1	26.1	78.3
	4	5	21.7	21.7	100.0
	Total	23	100.0	100.0	

According to table 3.14.0), 52.2% respondents disagreed that National entry tests have little or no contribution to the improvement of English language skills in Pakistanis, they consider its contribution as drastically below average.

(Table:3.14.1) Do you think that NTS Tests are merely a formality & do not focus student's university/college English language needs?

NEM2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	4	17.4	17.4	17.4
	3	6	26.1	26.1	43.5
	4	8	34.8	34.8	78.3
	5	5	21.7	21.7	100.0
	Total	23	100.0	100.0	

According to Table 3.14.1), 78% respondents considered NTS as merely a formality; its test patterns are not following the university/college English language criteria. Thus the respondents reported that they face many difficulties in teaching English language to Pakistani students at university/college level.

(Table:3.14.2) Do you think that English language test pattern of NTS is not focusing improve their creative writing ability?

NEM3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	2	8.7	8.7	17.4
	3	2	8.7	8.7	26.1
	4	11	47.8	47.8	73.9
	5	5	21.7	21.7	95.7
	10	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.14.2) reveals that73.9% respondents reported that the complete NTS English language test is objective and so it does not assess the pupil's creative writing skills, which are required for their academic and professional career. Therefore NTS is failing in meeting up the futuristic demands of academia in Pakistan.

(Table:3.14.3) Do you think that English language test pattern of NTS is not sufficient to prepare them for any International language proficiency tests likewise: IELTS, TOEFL,GMAT or GRE?

NEM4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	5	21.7	21.7	26.1
	3	2	8.7	8.7	34.8
	4	11	47.8	47.8	82.6
	5	3	13.0	13.0	95.7
	11	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

According to Table: 3.14.3), 82.6% respondents were of this opinion that there is no comparison between local entry tests standards and International tests standards. The chief complaint made by the respondents during this survey was that the students could not prepare themselves accordingly because NTS does not provide sound guidelines about its marking strategy and also lacks marking descriptors which are vital to measure students' proficiency according to standardized scales as is undertaken in the case of International examinations.

(Table:3.14.4) Do you think that local English language test would be beneficial for Pakistani students in future?

NEM5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	13.0	13.0	13.0
	3	5	21.7	21.7	34.8
	4	8	34.8	34.8	69.6
	5	6	26.1	26.1	95.7
	24	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.14.4 demonstrates that 69.6% respondents gave their opinion that if the NTS would be developed in terms of their exam strategies and evolve modern technology then it would prove to be a transparent and vital entry tests system in Pakistan.

Data set 2: Research objective 2:
2. To highlight the non-descriptive and irrational marking strategies both at formative and summative levels in Pakistan. (SPSS Data Code:(FSM)

Question cluster under objective 2. Code: FSM.

12. Do you think that English language courses in your university curriculum really fulfill student's academic needs?

15.). Are you satisfied with the grades your students get in your English language assessments in class?

16. They do not get good grades in English language tests it is because of their poor previous English language background?

17. They do not get good grades because you think curriculum scheme is boring?

18. They do not get good grades because you think that they lack skills required in question attempting?

19. They do not get good grades because you find that multiple choice questions are difficult for them?

20. They do not get good grades because they are not provided with frequent practice?

21. They do not get good grades because they lack effective reading & writing skills?

(Table:3.15.0) Do you think that English language courses in your university curriculum really fulfill student's academic needs?

FSM1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	2	8.7	8.7	17.4
	3	6	26.1	26.1	43.5
	4	10	43.5	43.5	87.0
	5	2	8.7	8.7	95.7
	12	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

According to table:3.15.0,, 87% respondents mutually agreed that English language courses fulfill the elementary requirements of university/college; they believed that Functional English, English comprehension & composition and critical writing skills help towards improving the students' exposure towards English as a second language.

(Table:3.15.1)). Are you satisfied with the grades your students get in your English language assessments in class?

FSM2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	13.0	13.0	13.0
	2	5	21.7	21.7	34.8
	3	4	17.4	17.4	52.2
	4	8	34.8	34.8	87.0
	5	2	8.7	8.7	95.7
	15	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

(Table:3.15.1) indicates that 87% respondents are satisfied with the grades their students acquire in English language assessments. Furthermore, the respondents reported that students having poor language background need to make an extra effort; thus students should be provided ample opportunities to improve while attempting multiple language quizzes and assignments during the semesters.

(Table:3.15.2) They do not get good grades in English language tests it is because of their poor previous English language background?

FSM3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	1	4.3	4.3	8.7
	3	5	21.7	21.7	30.4
	4	11	47.8	47.8	78.3
	5	4	17.4	17.4	95.7
	16	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

(Table:3.15.2) highlights the major concern of 78.3% respondents that English language teachers in Pakistan are not satisfied by the educational standards due to duel medium of instructions. The 78.3% respondents considered dual medium of instructions as an ultimate cause of students poor academic background. They have emphasized that mushroom growth of even so called English medium schools are not providing standardized instructional base.

(Table:3.15.3) They do not get good grades because you think curriculum scheme is boring?

FSM4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	5	21.7	21.7	30.4
	3	8	34.8	34.8	65.2
	4	6	26.1	26.1	91.3
	5	1	4.3	4.3	95.7
	17	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.15.3 illustrates that 65.2% respondents considered the English language course scheme as drab because the necessary and frequent revisions are not made by the Ministry of Education. Most of the respondents complain that the English language teachers are never taken into confidence to update the English language syllabi at university/college level.

(Table:3.15.4) They do not get good grades because you think that they lack skills required in question attempting?

FSM5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	4.3	4.3	4.3
	3	4	17.4	17.4	21.7
	4	13	56.5	56.5	78.3
	5	4	17.4	17.4	95.7
	18	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

According to Table:3.15.4, 78.3% respondents reported that students do not follow the instructional techniques of attempting questions; they were of this opinion that students are used to attempt questions in a traditional manner. They reported that most of the students attempt the questions in a superficial manner and they prefer to read the longer text rather than going through the questions at the first hand.

(Table:3.15.5) They do not get good grades because you find that multiple choice questions are difficult for them?

FSM6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	5	21.7	21.7	26.1
	3	6	26.1	26.1	52.2
	4	6	26.1	26.1	78.3
	5	4	17.4	17.4	95.7
	19	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.15.5 illustrates that 78.3% respondents were of this opinion that their students find multiple choice questions as difficult because they lack strategies to attempt the MCQS; 52.2% respondents gave partial response that at times MCQS are difficult for the students, while most of the time students get an advantage of MCQS, when attempting the options as a guess.

(Table:3.15.6)).They do not get good grades because they are not provided with frequent practice?

FSM7

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	2	8.7	8.7	13.0
	3	2	8.7	8.7	21.7
	4	11	47.8	47.8	69.6
	5	6	26.1	26.1	95.7
	20	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.15.6 indicates that 69.6% respondents were of this opinion that most of the students have not been provided frequent practice, due to which they could have had exposure to multiple variations of questions. Moreover, English language teachers have more than sufficient work load; therefore they cannot provide adequate practice sessions to the students in English language classrooms.

(Table:3.15.7) They do not get good grades because they lack effective reading & writing skills?

FSM8

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	2	8.7	8.7	8.7
	4	9	39.1	39.1	47.8
	5	11	47.8	47.8	95.7
	21	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.15.7 highlights that 95.7% respondents believe that Pakistani students are quite deficient in reading and writing skills. Furthermore, the respondents reported that many language students are reluctant to make extra effort in improving their reading and writing because they frequently complain that they have to cope with a large number of assignments related to their major technical subjects, so they cannot allocate enough time to reading and writing English texts.

3.16.0 Data set 3: Research objective 3:

3. To identify the inefficacy of National local testing bodies, which is a cause of de-motivation amongst Pakistani students. (SPSS Data Code:(INT).

Question cluster under objective 3. Code: INT.

3). Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?
4). If you disagree than do you think that NTS is a reliable & transparent assessment system?

5). Do you think that your students justify the score they acquired in NTS?

13). Do you think that due to NTS entry tests, it is difficult for students to obtain admission in universities?

14). Do you think that every university must conduct its own English language entry test?

22). Do you think that NTS entry test language portion is inadequate for students because it does not asses their listening and speaking skills?

(Table:3.16.0) Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?

INT1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	10	43.5	43.5	52.2
	3	6	26.1	26.1	78.3
	4	4	17.4	17.4	95.7
	5	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.16.0 indicates that 52.2% respondents are in favor that the universities/colleges must be given an authority to conduct their own aptitude tests because in this way they can assess the caliber of candidates accordingly their set standards, whereas 78.3% candidates were partially in favor of endowing test taking autonomy to the institutions. These respondents were of this opinion that partial autonomy can be given to the institutions so that every institution can implement its entry criteria effectively while reducing favoritism and encouraging meritocracy.

(Table:3.16.1) If you disagree than do you think that NTS is a reliable & transparent assessment system?

INT2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	2	8.7	8.7	17.4
	3	9	39.1	39.1	56.5
	4	10	43.5	43.5	100.0
	Total	23	100.0	100.0	

Table:3.16.1 illustrates that in case of disagreement 56.5% respondents remained neutral in their opinion about the reliability and transparency of NTS; these respondents were of this opinion that NTS can be reliable, if it improves its test patterns and rating strategies.

(Table:3.16.2) Do you think that your students justify the score they acquired in NTS?

INT3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	5	21.7	21.7	30.4
	3	7	30.4	30.4	60.9
	4	7	30.4	30.4	91.3
	5	2	8.7	8.7	100.0
	Total	23	100.0	100.0	

Table:3.16.2 indicates that 91.3% respondents claimed that their students justify the score which they acquire at the entry level assessment. At the same time, 60.9% respondents gave partial opinion because they think that their students English language background is too weak to justify even average scores and so the learners do not meet the standard English language criteria at the university level.

(Table:3.16.3) Do you think that due to NTS entry tests, it is difficult for students to obtain admission in universities?

INT4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	30.4	30.4	30.4
	2	5	21.7	21.7	52.2
	3	5	21.7	21.7	73.9
	4	5	21.7	21.7	95.7
	13	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.16.3 indicates that 95.7% respondents considered NTS as a barrier in acquiring admissions in universities because many students underestimate themselves and give up their efforts for continuing education. This might be responsible for increased illiteracy in Pakistan. 73.9% respondents remained partial because they were of this opinion that there must be sound filtration at the entry level admissions.

(Table:3.16.4) Do you think that every university must conduct its own English language entry test?

INT5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	3	13.0	13.0	17.4
	3	2	8.7	8.7	26.1
	4	12	52.2	52.2	78.3
	5	4	17.4	17.4	95.7
	14	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.16.4 indicates that 78.3% respondents were of this opinion that there must be separate English language tests at the entry level so that students would get ample opportunity to prepare themselves to meet the International English language proficiency criteria, while preparing themselves for the local English language test criteria in Pakistan.

(Table:3.16.5) Do you think that NTS entry test language portion is inadequate for students because it does not asses their listening and speaking skills?

INT6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1	4.3	4.3	4.3
	2	2	8.7	8.7	13.0
	3	3	13.0	13.0	26.1
	4	10	43.5	43.5	69.6
	5	6	26.1	26.1	95.7
	22	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.16.5 shows that 69.6% respondents considered NTS as a non-standard English language assessment test because it does not assess the pupils' listening and speaking skills, which are considered important part of the criteria determining admission in a well reputed university both at the national and international level. Moreover, in order to acquire high profile jobs the young students may be required to demon proficient spoken and listening abilities so that they meet their career needs as well.

3.17.0 Data set 4: Research objective 4:

4.To identify the reasons of lack of unified local assessment criteria. (SPSS Data Code: (NAC).

Question cluster under objective 4. Code. (NAC)

6). You are not satisfied by their class performance because you think they are not properly assessed by NTS?

7). You are not satisfied because you think English language assessment was too difficult in NTS?

8). You are not satisfied because you think that their numerical ability is stronger than their your language ability?

9). Do you think that English language Multiple choice questions are sufficient to assess their English language proficiency?

23). Do you think that there should be a separate local English language test, as university entry criteria?

(Table:3.17.0) You are not satisfied by their class performance because you think they are not properly assessed by NTS?

NAC1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	30.4	30.4	30.4
	2	4	17.4	17.4	47.8
	3	2	8.7	8.7	56.5
	4	9	39.1	39.1	95.7
	6	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.17.0 illustrates that 95.7% respondents agreed that due to lack of sound summative benchmarks, students' critical writing skills are not assessed at entry level. Therefore, students generally fail to meet the expected standards of the university. Due to this ineffective summative evaluation, student's failure rate is higher in English language tests as compared to other compulsory subjects at the university level.

(Table:3.17.1) You are not satisfied because you think English language assessment was too difficult in NTS?

NAC2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	21.7	21.7	21.7
	2	7	30.4	30.4	52.2
	3	7	30.4	30.4	82.6
	4	3	13.0	13.0	95.7
	7	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.17.1 indicates the neutrality of 82.6% respondents; most of the respondents reported that they do not think that NTS English language assessment is too difficult. At the same time, the respondents were of the opinion that it depends on the exposure and English language background of the candidates. They also believed that the Pakistani students display diverse English language backgrounds.

(Table:3.17.2)You are not satisfied because you think that their numerical ability is stronger than their your language ability?

NAC3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.7	8.7	8.7
	2	3	13.0	13.0	21.7
	3	5	21.7	21.7	43.5
	4	11	47.8	47.8	91.3
	5	1	4.3	4.3	95.7
	8	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.17.2 shows that 91.3% respondents agreed that Mathematics is the favorite subject of most of the engineering students, whereas English language proves to be quite confusing for them because they cannot apply basic grammatical rules. Therefore they lack creative writing skills to a great extent.

(Table:3.17.3) 9). Do you think that English language Multiple choice questions are sufficient to assess their English language proficiency?

NAC4

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	4	17.4	17.4	17.4
	2	7	30.4	30.4	47.8
	3	2	8.7	8.7	56.5
	4	8	34.8	34.8	91.3
	5	1	4.3	4.3	95.7
	9	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

According to Table:3.17.3, 91.3% respondents believed that the multiple choice questions are sufficient to assess the basics of English language of mediocre students. They were of the opinion that the majority of the Urdu medium and even English medium students have extremely poor English language background; therefore MCQs help the English language teachers to identify the students' basic English language background to a great extent.

(Table:3.17.4). Do you think that there should be a separate local English language test, as university entry criteria?

NAC5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	1	4.3	4.3	4.3
	3	4	17.4	17.4	21.7
	4	11	47.8	47.8	69.6
	5	6	26.1	26.1	95.7
	23	1	4.3	4.3	100.0
	Total	23	100.0	100.0	

Table:3.17.4 indicates that 69.6% respondents agreed that there should be a separate English language test conducted at the entry level, so that the candidates may be bound to really work hard to acquire proficiency in English. Moreover, it will provide ample opportunity to the candidates to improve their English language; consequently, the students may better cope with their academic needs and also they can become sufficiently proficient to compete at the international level.

Chapter 4:

4.0 Results Summary:

Addressing to the research question 1, the findings of this research show that many examiners serving National testing system in Pakistan are educationists but they are not professional examiners therefore, they are not familiar with the formulation of test items. They have not acquired specific rater training, moreover, they are not familiar with test item constructions, as a matter of fact plagiarism is reported frequently by many educational institutions. In addressing the research question 2, the findings of this research show that many English language teachers reported that they have not acquired in-service or pre-service training in the application of marking descriptors, because the English language teachers in Pakistan lack both the exposure and research in utilizing the international standardized marking mechanism.

In addressing to the research question 3, the research findings show that diverse educational backgrounds and confused dual mediums of instruction (Urdu and English) and also lack of investment on massive research in the improvement of testing and evaluation in Pakistan are primary reasons of the ineffective local examination systems.

In addressing to the research question 4. The research findings show that lack of specific benchmarks, absence of quality assurance, untrained examiners, and lack of frequent monitoring are the major factors effecting the transparency of local assessment criteria in Pakistani examination systems.

Students' Survey Results:

4.1Results of SPSS Data Code NEP:

93% respondents are in favor that if the National testing system in Pakistan gets more scrutinized than it can become a vital source of students improvement in English language skills in Pakistan. 88% respondents are of this opinion that National testing system in Pakistan is merely a formality and it has nothing to do with their basic academic requirements in university or college because the formative level of English language assessments is comparatively higher than the local level assessments.

87% respondents were of this opinion that the NTS test pattern in case of English language assessments has no contribution in improving their creative writing skills because the examination format is more objective.74% respondents showed their dissatisfaction with the standard of the National Testing System in comparison to the International examination systems. Therefore, the local testing bodies are not considered as benchmarks by the Pakistani students.

73% respondents desired improvement in the NTS test patterns; according to the respondents, it can become a sound benchmark in future, if the test patterns are designed according to the International standards.

4.2Results of SPSS Data Code: (FSM) :

87% respondents showed their satisfaction about the standards of English language syllabi prescribed in their universities; they believed that the syllabus is meeting the academic requirements of the respondents, and through the prescribed scheme of study they can easily compete at the International level. 81% respondents are satisfied by the grades they get in the subject of English language at university level. They believed that they really need to work hard in order to have satisfactory grades; furthermore, they responded that the English

language course in their universities is really effective in improving their background knowledge.

85% respondents agreed that they have poor academic background due to insufficient instructional base. They also showed their dissatisfaction from the so called English medium schools and also the dual standards likewise: Urdu & English Mediums 92% respondents from Urdu medium background were of the opinion that the prescribed curriculum has no innovation, the syllabus needs frequent revisions. 84% respondents agreed that they lack question attempting skills, therefore, they have poor grades in English Language according to the university/college assessments.82% agreed that they really have had little exposure to the multiple choice questions attempting techniques, which is considered an important part of English language objective tests both in the entry level tests and in-house examinations.

The respondents admitted their own negligence for not preparing well; they considered lack of appropriate and frequent practice as the major cause of their low grades in English language tests.89% respondents admitted that the cause of their poor proficiency is the lack of good reading and writing skills; they admitted their reluctance in improving their language proficiency skills.

4.3Results of SPSS Data Code: (INT):

94% respondents strongly agreed that the entry test autonomy must not be given to the universities/colleges because they might set their own biased criteria, which may lead to unfairness in assessment. The respondents strongly preferred improvements in the NTS as an objective examination body.86% respondents are in favor of an unanimous and transparent examination body which should be highly mechanized and must have a central authority as far as admissions are concerned. 88% respondents agreed that their NTS scores were justified to a great extent.79% respondents consider NTS as neither difficult nor easy; the respondents were of this opinion that all parts of the question paper are interrelated therefore, Numerical attempt is equally required to be scored well as English language portion in NTS.

85% respondents are of this opinion that the universities/colleges must conduct an English language entry test which is beneficial for the students and also the institutional authorities must have an idea about the strengths and weaknesses of the pupil they are inducting. 76% respondents had reported that listening & speaking skills are not assessed in any local entry test in Pakistan; therefore the pupils remain ignorant about their linguistic deficiencies in this respect and they also suffer in future due to these linguistic in competencies.

4.4Results of SPSS Data Code: (NAC)

The 82.6% respondents were unanimously agreed that due to the absence of unified assessment criteria, the failure rate is high, students feel English language assessment part in NTS test as challenging, on the other hand 91.3% respondents were of this opinion that the Multiple choice questions provide majority of the students an ease in attempting questions in the local entry tests. 69.6 % respondents suggested that there should be a separate English language test at the entry level, this will improve the linguistic deficiencies of the students and later on it will help them to cope the academic challenges at the university level.

95.7% respondents agreed that due to lack of sound summative benchmarks, students' critical writing skills are not assessed at entry level. Therefore, students generally fail to meet the expected standards of the university. Due to this ineffective summative evaluation, student's

failure rate is higher in English language tests as compared to other compulsory subjects at the university level.82.6% respondents; most of the respondents reported that they do not think that NTS English language assessment is too difficult. At the same time, the respondents were of the opinion that it depends on the exposure and English language background of the candidates. They also believed that the Pakistani students display diverse English language backgrounds.

91.3% respondents agreed that Mathematics is the favorite subject of most of the engineering students, whereas English language proves to be quite confusing for
them because they cannot apply basic grammatical rules. Therefore they lack creative writing skills to a great extent.91.3% respondents believed that the multiple choice questions are sufficient to assess the basics of English language of mediocre students. They were of the opinion that the majority of the Urdu medium and even English medium students have extremely poor English language background; therefore MCQs help the English language teachers to identify the students' basic English language background to a great extent.

69.6% respondents agreed that there should be a separate English language test conducted at the entry level, so that the candidates may be bound to really work hard to acquire proficiency in English. Moreover, it will provide ample opportunity to the candidates to improve their English language; consequently, the students may better cope with their academic needs and also they can become sufficiently proficient to compete at the international level.

Teachers' Survey Results:

4.5Results of Data code: NEP:
93% respondents are in favor that if the National testing system in Pakistan gets more scrutinized than it can become a vital source of students improvement in English language skills in Pakistan.88% respondents are of this opinion that National testing system in Pakistan is merely a formality and it has nothing to do with their basic academic requirements in university or college because the formative level of English language assessments is comparatively higher than the local level assessments.

87% respondents were of this opinion that the NTS test pattern in case of English language assessments has no contribution in improving their creative writing skills because the examination format is more objective.74% respondents showed their dissatisfaction with the standard of the National Testing System in comparison to the International examination systems. Therefore, the local testing bodies are not considered as benchmarks by the Pakistani students.73% respondents desired improvement in the NTS test patterns; according to the respondents, it can become a sound benchmark in future, if the test patterns are designed according to the International standards.

4.6Results of SPSS Data Code: (FSM)
87% respondents mutually agreed that English language courses fulfill the elementary requirements of university/college; they believed that Functional English, English comprehension & composition and critical writing skills help towards improving the students' exposure towards English as a second language. 87% respondents are satisfied with the grades their students acquire in English language assessments. Furthermore, the respondents reported that students having poor language background need to make an extra

effort; thus students should be provided ample opportunities to improve while attempting multiple language quizzes and assignments during the semesters.

78.3% respondents that English language teachers in Pakistan are not satisfied by the educational standards due to duel medium of instructions. The 78.3% respondents considered dual medium of instructions as an ultimate cause of students poor academic background. They have emphasized that mushroom growth of even so called English medium schools are not providing standardized instructional base.

65.2% respondents considered the English language course scheme as drab because the necessary and frequent revisions are not made by the Ministry of Education. Most of the respondents complain that the English language teachers are never taken into confidence to update the English language syllabi at university/college level.

78.3% respondents reported that students do not follow the instructional techniques of attempting questions; they were of this opinion that students are used to attempt questions in a traditional manner. They reported that most of the students attempt the questions in a superficial manner and they prefer to read the longer text rather than going through the questions at the first hand. 78.3% respondents were of this opinion that their students find multiple choice questions as difficult because they lack strategies to attempt the MCQS;

52.2% respondents gave partial response that at times MCQS are difficult for the students, while most of the time students get an advantage of MCQS, when attempting the options as a guess.

69.6% respondents were of this opinion that most of the students have not been provided frequent practice, due to which they could have had exposure to multiple variations of questions. Moreover, English language teachers have more than sufficient work load; therefore they cannot provide adequate practice sessions to the students in English language classrooms.

95.7% respondents believe that Pakistani students are quite deficient in reading and writing skills. Furthermore, the respondents reported that many language students are reluctant to make extra effort in improving their reading and writing because they frequently complain that they have to cope with a large number of assignments related to their major technical subjects, so they cannot allocate enough time to reading and writing English texts.

4.7Results of SPSS Data Code: (INT)

52.2% respondents are in favor that the universities/colleges must be given an authority to conduct their own aptitude tests because in this way they can assess the caliber of candidates. Accordingly to the set standards, whereas

78.3% candidates were partially in favor of endowing test taking autonomy to the institutions. These respondents were of this opinion that partial autonomy can be given to the institutions so that every institution can implement its entry criteria effectively while reducing favoritism and encouraging meritocracy.

56.5% respondents remained neutral in their opinion about the reliability and transparency of NTS; these respondents were of this opinion that NTS can be reliable, if it improves its test patterns and rating strategies.

91.3% respondents claimed that their students justify the score which they acquire at the entry level assessment. At the same time, 60.9% respondents gave partial opinion because they think that the Pakistani students' English language background is too weak to justify even average scores and so the learners do not meet the standard English language criteria at the university level.

78.3% respondents were of this opinion that there must be separate English language tests at the entry level so that students would get ample opportunity to prepare themselves to meet the International English language proficiency criteria, while preparing themselves for the local English language test criteria in Pakistan.

69.6% respondents considered NTS as a non-standard English language assessment test because it does not assess the pupils' listening and speaking skills, which are considered important part of the criteria determining admission in a well reputed university both at the national and international level. Moreover, in order to acquire high profile jobs the young students may be required to demonstrate proficient spoken and listening abilities so that they could meet their career needs as well.

4.8 Results of SPSS Data Code: (NAC):

The 82.6% respondents were unanimously agreed that due to the absence of unified assessment criteria, the failure rate is high, students feel English language assessment part in NTS test as challenging, on the other hand

91.3% respondents were of this opinion that the Multiple choice questions provide majority of the students an ease in attempting questions in the local entry tests.

69.6 % respondents suggested that there should be a separate English language test at the entry level, this will improve the linguistic deficiencies of the students and later on it will help them to cope the academic challenges at the university level.

4.9 Pearson Correlation Results:

The results of this research were gathered through intensive planning while utilizing the statistical tools like Pearson Correlations through the statistical software SPSS 17.0.

Brief introduction of Pearson Correlation Significance:

A statistically *significant* finding is one that is determined (statistically) to be very unlikely to happen by chance. Statisticians are able to calculate the likelihood that any observed relationship between two variables (as indicated by any number of cases) could have happened by chance (or random variation). If it is calculated that there is less than one in twenty chance (.05 or 5%) that the observed relationship could have happened by chance, the findings are designated as significant. If there is less than one in one hundred chance (.01 or 1%), they are designated as highly significance. Significance is influenced by the number of cases in your sample, and the observed range (difference) of the sample. Simply put, the more likely to be sure that the differences observed from a sample are accurate for the whole population if there are many cases and large comparative differences in the observed relationship between a specific set of variables. This text indicates significance by placing one or two asterisks (*) after the Pearson's r.

Table:4.9

Correlations

		NEM	INT	NAC	FSM
NEM	Pearson Correlation	1			
	Sig. (2-tailed)				
	N	100			
INT	Pearson Correlation	.252*	1		
	Sig. (2-tailed)	.011			
	N	100	100		
NAC	Pearson Correlation	.199*	.369**	1	
	Sig. (2-tailed)	.048	.000		
	N	100	100	100	
FSM	Pearson Correlation	.231*	.270**	.382**	1
	Sig. (2-tailed)	.021	.007	.000	
	N	100	100	100	100

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

**. Correlation is significant at the 0.01 level (2-tailed).

4.10 Correlation Results: The above stated table reveal strong relation between this research variables. According to the above stated correlation data the results are significant and the issues defined in this research variables require special attention. This correlation data makes this research more authentic.

Chapter 5.

5.0 Conclusion:

The above stated summary indicates that the respondents demand the existence of transparent examination systems, the research variables are depicting a strong relation, which indicates that the research hypothesis is justified, and the local examination system really needs to be improved, the local testing bodies in Pakistan have their crucial role to play. This research data reveals that respondents showed their distrust on local examining bodies like: NTS. It is essential that the policy makers must pay attention to this highly ignored aspect and devise some sound think tank.

The examination system of any country is the key to the future prospects of the learners. On the other, hand this research highly emphasizes that summative and formative assessments are the part and parcel to each other, whatever, the young learners acquire in their schools, it effects their output when they appear in some entry test.

In addition to curricular and instructional alignment, assessments should provide information as to where students are along an identified learning progression so that teachers can plan appropriate next instructional steps (Heritage, 2008). Understanding how the underlying skills and knowledge of a given instructional unit connect is critical to a teacher's ability to evaluate individual student responses. A misconception presented early in the learning process may be addressed quite differently than if it were to appear later in the instructional unit. An inherent need within sound learning progressions is a thorough understanding of the short and long-term intended learning objectives.

That the teacher should understand these objectives are fundamental to a sound curriculum (Tyler, 1949), however in the context of formative processes the understanding of the learning objectives must extend to the students. While cognitive theories describe the importance of met cognition in the process of learning, students need to know the intended outcome and how their work will be judged as they consider their own mental strategies (i.e. thinking about their thinking) in solving a given problem. For example, Artery (2000) described two equal purposes for scoring rubrics:

a) as a tool for teachers to evaluate and track student progress.

b) as a tool for students to improve performance against a known criterion.

Teachers should attend to both of these purposes as they use a given rubric. The necessary involvement of students in the understanding of learning objectives and evaluation criteria requires a "shift" from teachers being primarily responsible for student learning to a classroom context in which students "assume meaningful responsibility for their own learning and the learning of their classmates" (Popham, 2008, pp. 94-95).

The notion that assessment information should sit within, rather than apart from, the learning process is also an important facet of context. The well-intentioned triangle we have often seen depicting the interconnectedness of curriculum, assessment, and instruction may better describe a summative process that treats these elements as related, yet separate. A formative system should view ongoing assessment within a learning process. This shift in the "learning culture" is what Sheppard (2000) presented as needed "…so that students and teachers look to assessment as source of insight and help instead of an occasion for meeting out rewards and punishments".

.Specifically, Black and William's comprehensive meta-analysis (1998a) and subsequent summary (1998b), Sheppard's (2000) conceptualization of classroom assessment within the context of cognitive, learning, and curricular theories, and NRC's How People L*earn* (2000) and *Knowing What Students Know* (2001) collectively noted the importance of:

- Providing quality, descriptive feedback;
- Using effective questioning techniques;
- Assessing prior knowledge
- Misconceptions.
- Implementing student goal setting,
- Self-regulation, and self-evaluation.

While keeping the current deteriorating examination standards in Pakistan, this study leads to the below given recommendations which can be proved beneficial in order to meet the futuristic demands of testing and evaluation systems in Pakistan.

5.1 Recommendations:

- The Local Pakistani universities must encourage their English language teachers to conduct wide scale research on the English language testing methods; they must be encouraged to work in collaboration with NTS.

- An English language criteria assessment board can also be developed to monitor the test materials both at the NTS and at the university levels to ensure quality English language Testing mechanisms. Moreover, external board can also be formulated in collaboration with foreign testing bodies.

- CBT (Computer based testing) must also be introduced to share the examiners marking load; it can be helpful to assess MCQs and other objective type questions, so that English language examiners are not reluctant to add subjective portions in English language tests. Therefore in this way the creative and critical writing skills can also be given weight age in the local entry tests.

- A think tank must be devised to design the marking descriptors which are based on several sets of proficiency scales; in this way the candidates will be more familiar with the examination strategies and can prepare accordingly for the test.

- The foreign test taking agencies and examiners must be involved in devising and constructing the test batteries which can be beneficial for the Pakistani examiners; in this way transparent entry test mechanism can be developed in Pakistan.

- A wide scale English language immersion program can be introduced in Pakistan and a collaborative teacher training program can also be conducted. Internationally accepted examination bodies like IDP(International development programs for universities and colleges, Australia), British council and so on can also be involved.

- The above discussion also reinforces the fact that we need to energize the curriculum. A curriculum should be designed that blends thinking skills with the affective domain; information, media, ICT literacy, life and career skills in the context of core academic subjects and cross interdisciplinary themes.

- Students must be provided with opportunities to engage in virtual tours, web networks and generic and subject specific interactive softwares to provide meaningful, relevant, authentic learning within and beyond the classrooms.

- In order to assure that the Pakistani examination system meets certain standards of quality control. The tasks need to be moderated by an editing committee (a group of experienced item writers), then piloted with real students, and finally revised in the light of the pilot examinations. It usually takes several rounds of careful editing, piloting and revision before a task is given its final shape.

- Writing is a subjective test. Examiners of writing skills are not making a simple decision whether an answer is right or wrong but have to do something more complex: they can use a rating scale consisting of several bands and descriptors to assess a script. It is very important that examiners understand what rating scales measure and that they can interpret correctly and use consistently the descriptors in the scales.

- Two things are necessary in order to be able to do this: (1) the rating scale used for the assessment of writing must be valid and reliable, and (2) the examiners must be reliable. A reliable and valid rating scale is the result of a long process during which the rating scale is tried out in order to test, for instance, whether raters interpret the descriptors the same way, whether the scale is suitable to make fine distinctions between different quality scripts, or whether it is appropriate forth assessment of the writing tasks to which it is applied. If the rating scale is suitable for rating writing, the next step is to ensure that it is used appropriately. All the work invested in designing a writing scale can be wasted if the scale is not used correctly and consistently by all the examiners. Obviously, the more examiners use a scale the more likely it is that it will not be used consistently unless all the examiners have received training in its use.

- However carefully designed and easy-to-use a scale is, a group of untrained markers using it will most probably give different marks for the same script. It does not matter whether the examiner is a writing expert or a language teacher who does not deal more with writing than teach the odd writing unit from a course book. Experts or not, all examiners will interpret the scale their own way and this will result in unreliable assessment. Without training it may happen, for example, that the same script gets two different scores, maybe one with which the candidate fails and one with which he/she passes the exam, from two examiners working in two different examination centres.

- In order to make sure that the rating is reliable in NTS, and that examiners can be trained, standards must be set. In writing assessment this is done with benchmarking. Bench-marking is a process that must be repeated every time before the rating begins, especially in the case of an examination that is administered rarely, for example, once a year. After the examination, a chief examiner and a small group of experts read a number of the scripts produced in the examination, and keeping the rating scale in

mind select some that represent various levels of performances (i.e. weak, average, good scripts). Once the scripts are selected, the chief examiner and the group of experts, that is the members of the benchmarking committee, rate each script individually, using the rating scale. When they have rated the scripts, the members of the bench-marking committee meet to compare and discuss the marks they have awarded to the scripts. Their aims are to reach a consensus concerning the marks awarded to the scripts and if necessary to improve the rating scale.

- The marks that are agreed upon by the members of the bench-marking committee for the scripts are called benchmarks. The benchmarked scripts represent different levels of performance and are used as model performances with which examiners can be trained in what is called a standardization meeting or rater training. The participants in the rater training are the chief examiner or another member of the bench-marking committee and a number of examiners. The aim of the meeting is to help examiners understand the rating scale and the rating process. Before the rater training examiners rate individually some of the scripts selected by the bench-marking committee, and make a note of the reasons why they awarded a particular score. At the meeting the benchmarks are compared with the marks given by the examiners. If there are differences, the examiners must be given the reasons that justify the experts' scores. The aim of this meeting is to help the examiners to learn to use the rating scale in the same way as the experts.

- In the second part of the rater training, the examiners rate individually the remaining benchmarked scripts and this time their scores are expected to match the benchmarks. If their scores match the benchmarks, it means that the examiners have learnt to interpret the
- descriptors of the rating scale the same way as the members of the bench-marking committee. This in turn is expected to mean that two examiners will award the same score or almost identical scores for the same script when they mark the rest of the scripts. However, even if they have participated in the training program, it may happen that because of various reasons (e.g. place and time of rating) the examiners occasionally fail to follow the standards they learned during the rater training. This is especially dangerous if the scripts are marked by one examiner only. Modern European examinations therefore use double marking in writing assessment. This means that each script is rated independently by two examiners. If the two examiners have applied the rating scale in accordance with the practice developed in the course of the rater training, they will award the same or closely similar scores for the same script. Different examinations have different procedures for the cases when the scores of two examiners differ. For example, if the difference is small, the average of the two scores may be calculated. If the difference is big, a third rater is asked to rate the script independently.

- In order to check whether the examiners apply the rating scale consistently and correctly, their intra-rater and inter-rater reliability can be checked. The intra-rater reliability check shows whether an examiner agrees with him or herself when rating the same script on different occasions. Ideally, if an examiner rates a script on two different days, he or she should award the same score. If this happens, it means that they interpreted the rating scale in the same way on both occasions. However, even if they agree with themselves, examiners may not agree with other raters who use the same scale to rate the same script. They may be consistent but idiosyncratic and as a consequence unreliable. The inter-rater reliability check shows whether two different

examiners apply the rating scale in the same way, that is, whether there is agreement between the raters. Two examiners rating the same script are expected to give the same or closely similar scores.

- In order to ensure that the candidates are assessed fairly and that they do not fail their examination because of the unreliability of an examiner, raters should use reliable and valid scales. They should regularly participate in a training program to learn to use standards consistently. Test administrators should use double marking and apply routine checks to find those examiners whose rating is unreliable. This is especially important in the case of a high-stakes examination like a school-leaving or an entrance examination likewise NTS tests.

References:

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). ***Standards for educational and psychological testing.*** Washington, DC: American Educational Research Association.

Arter, J. (2000, April). ***Rubrics, scoring guides, and performance criteria: Classroom tools for assessing and improving student learning***. Paper presented at an annual meeting of the American Educational Research Association, New Orleans, LA.

Black, P. & Wiliam, D. (1998a). Assessment and classroom earning. ***Assessment in Education: Principles, Policy, & Practice. 5***(1), 7-74.

Black, P. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. ***Phi Delta Kappan. 80***(2), 139-144.

Goertz, M. Olàh, & Riggan, M. (2009, Dec.). From testing to teaching: The use of interim assessments in classroom instruction. Consortium for Policy Research in Education (CPRE) Research Report #RR-65. Retrieved on May 15, 2010, from http://CPRE.org

Heritage, M. (2008). Learning progressions: Supporting instruction and formative assessment. Council of Chief State School Officers: Washington, DC. Retrieved on August 20, 2010, from http://www.ccsso.org/Resources/Publications/Learning_Progressions_Supporting_Instruction_and_Formative_Assessment.html

Practical Assessment, Research & Evaluation, Vol 16, No 3 Page 6 Good, Formative Use of Assessment Information Heritage, M., Kim, J., Vendlinski, T.P., & Herman, J.L. (2008). ***From evidence to action: A seamless process in formative assessment?*** (CRESST Report 741). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

McManus, S. (2008). Attributes of effective formative assessment. Council of Chief State School Officers: Washington, DC. Retrieved on July 19, 2010, from http://www.ccsso.org/Resources/Publications/Attributes_of_Effective_Formative_Assessment.html

Messick, S. (1989). Validity. In R.L. Linn (Ed.), ***Educational Measurement*** (3rd ed., pp. 13-103).

New York: American Council on Education and Macmillan Publishing Co. National Research Council (2000). ***How people learn: Brain, mind, experience, and school***. Washington, D.C.: National Academy Press.

National Research Council (2001). ***Knowing what students know.*** Washington, D.C.: National Academy Press. Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. ***Educational Measurement: Issues and Practice, 28***(3), 5-13.

Popham, W.J. (2008). ***Transformative assessment.*** Alexandria, VA: Association for Supervision and Curriculum Development.

Shepard, L. (2000). The role of assessment in learning Culture. ***Educational Researcher, 29***(7), 4-14.

Shepard, L. (2005, Oct.). Formative assessment: Caveat emptor. Paper presented to an Educational Testing Service Invitational Conference, New York.

Stiggins, R., Arter, J., Chappuis, J. and Chappuis, S. (2006). ***Classroom assessment for student learning: Doing it right – Using it well.*** Portland, OR: Educational Testing Service. Tyler, R. W. (1949). ***Basic principles of curriculum and instruction.*** Chicago: University of Chicago Press.

Wiliam, D. (2000, Nov.). ***Integrating summative and formative functions of assessment.*** Keynote address to the European Association for Educational Assessment, Prague, Czech Republic. Retrieved on Jan. 17, 2011 from http://eprints.ioe.ac.uk/1151/1/Wiliam2000IntergratingAEAE_2000_keynoteaddress.pdf

Vygotsky, L. S., (1978). ***Mind in society: The development of higher psychological processes.*** Cambridge, MA: Harvard UniversityPress.

(APA) American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards foreducational and psychological testing*. Washington: American Psychological Association.

Apple, M. (1982). *Education and power*. Boston: Routledge and Kegan Paul.Arendt, H. (1969). *On violence*. London:

Penguin.Australian National Training Authority. (1994). *Towards a skilled Australia: A national strategy for vocational education and training* : Australian National Training Authority.
Ayers, W. (1993). *To teach: The journey of a teacher*. New York: Teachers College Press.
Ball, S. (1994). *Education reform*. Buckingham: Open University Press.
Barone, T. (1992). Beyond theory and method: A case of critical storytelling. *Theory into Practice*, 31(2), 143-146.
Barton, L., Whitty, G., Miles, S., & Forlong, J. (1994). Teacher education and teacher professionalism in England: some emerging issues. *British Journal of Sociology in education, 15(4)*, 529-543.
Bateson, G. (1972). *Steps to an ecology of mind*. New York: Ballantine Books.
Bateson, G. (1979). *Mind and nature*. London: Wildwood House.
Becker, H. (1990). Generalising from case studies. In E. Eisner & A. Peshkin (Eds.) *Qualitative enquiry in education: the continuing debate*. New York: Teachers College, Columbia University.
Beevers, B. (1993). Competency-based training in TAFE: Rhetoric and reality. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.
Behar, I. (1983). *Achievement Testing*. Beverly Hills: Sage Publications.
Beittel, K. (1984). Great swamp fires I have known: Competence and the hermeneutics of qualitative experiencing. In E. Short (Ed.), *Competence: Inquiries into its meaning and acquisition in educational settings*, (pp. 105-122). Lanham, MD: University Press of America.
Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment and Evaluation in Higher Education, 18(2)*, 83-93.
Berk, R. (1986). A consumers guide to setting performance standards on criterion reference tests. *Review of Educational Research, 56(1)*, 137-172.
Biesta, G. (1994). Education as practical intersubjectivity: towards a critical-pragmatic understanding of education. *Educational Theory, 44(3)*, 299-317.
Bloom,B. (Ed.). (1956). *Taxonomy of educational objectives; Handbook 1, cognitive domain*. New York: David Mc Kay.
Bloom, B. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.
Bloom, B., Hastings, J., & Madaus, G. (1964). *Handbook on formative and ummative evaluation of student learning*. New York: McGraw Hill.
Borthwick, A. (1993). Key competencies - Uncovering the bridge between the general and vocational. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.
Bourdieu, P., & Passeron, J. (1977). *Reproduction in education, society and culture*. London: SAGE Publications.
Bowden, J., & Masters, G. (1993). *Implications for higher education of a competency-based approach to education and training*. Canberra: Australian Government Publishing Service.
Bracht, G., & Glass, G. (1968). The external validity of experiments. *American Educational Research Journal, 5(4)*, 437-474.
Broadfoot, P. (Ed.). (1984). *Selection, certification and control*. London: The Falmer Press.
Brown, R. (1973). *Religion and violence*. Philadelphia: The Westminster Press.
Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge: Harvard University Press.
Bucke, R. (1969). *Cosmic consciousness*. New York: E.P. Dutton Co.
Burchell, G., Gordon, C., & Miller, P. (Eds.). (1991). *The Foucault effect*. London: Harvester Wheatsheaf.
Burgess, R. (Ed.). (1985). *Issues in educational research: Qualitative methods*. London: The Falmer Press.
Burton, N. (1978). Societal standards. *Journal of Educational Measurement, 15(4)*, 263-273.
Cairns, L. (1992). Competency-based education: Nostradamus's nostrum. *The Journal of Teaching Practice, 12(1)*, 1-32.
Camera, H. (1971). *Spiral of violence*. London: Sheed and Ward.

Campbell, J. (1956). *Hero with a thousand faces*. New York: Meridian Books. Carr, W., & Kemmis, S. (1983). *Becoming critical*. Geelong: Deakin University Press.
Cherryholmes, C. (1988). *Power and criticism*. New York: Teachers College Press.
Cherryholmes, C. H. (1988). Construct validity and the discourses of research. *American Journal of Education, 96(3)*, 421-457.
Clough, E. E., Davis, P., & Sumner, R. (1984). *Assessing pupils: a study of policy and practice.* Windsor: NFER-Nelson.
Codd, J. (1985). *Curriculum discourse: text and context.* Paper presented at the National Conference of the Australian Curriculum Studies Association, La Trobe University, Melbourne.
Codd, J. (1988). The construction and deconstruction of educational policy documents. *Journal of Education Policy, 3(3),* 235-247.
Collins, C. (Ed.). (1993). *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.
Collins, R. (1979). *The credential society*. Orlando: Academic Press Inc. Cox, R., & 1965. (1965). *Examinations and higher education: A survey of the literature*. London: Society for Research into Higher Education. Cresswell, M. (1995). Technical and educational implications of using public examinations for selection to higher education. In T. Kellaghan (Ed.), *Admission to higher education*, . Dublin: Educational Research Centre.
Cronbach, L. (1969,). *Validation of educational measures.* Paper presented at the The 1969 invitational conference on testing problems: Towards a theory of achievement measurement.
Cronbach, L. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum. Cronbach, L., Rajaratman, N., & Gleser, G. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, XVI(2)*.
Cronbach, L. J. (1990). *Essentials of psychological testing*. (Fifth ed.). New York: Harper and Row.
Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: performance assessment. *Educational Researcher, 23(5),* 11-18.
Docking, R. (1995, January 1995). Competency: What it means and how you know it has been achieved. *NTB Network- Special Conference Edition, 18*.
Donmeyer, R. (1990). Generalizability and the general case study. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education*, . New York: Teachers College, Colombia University.
Downs, C. (1995). *Key competencies: A useful agent for change?* . Richmond: National Centre for Competency Based Assessment and Training.
Eisner, E. (1988). The primacy of experience and the politics of method. *Educational Researcher, 17(3),* 15-20.
Eisner, E. (1990). The meaning of alternative paradigms. In E. Guba (Ed.), *The paradigm dialog*, Newbury Park: Sage Publications.
Eisner, E. (1991b). Taking a second look: Educational connoisseurship revisited. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp.169-187). Chigago: The National Society for the Study of Education.
Eisner, E., & Peshkin, A. (Eds.). (1990). *Qualitative enquiry in education*.
Eisner, E. W. (1985). *The educational imagination*. (second ed.). New York: Macmillan.
Eisner, E. W. (1991). *The enlightened eye*. New York: Macmillan.
Fay, B. (1987). *Critical social science*. New York: Cornel University Press.
Feyerabend, P. (1988). *Against method*. London: Verso.
Finn, B. C. (1991). *Young people's participation in post-compulsory education and training* . Canberra: Australian Educational Council Review Committee.
Fish, S. (1980). *Is there a text in the class? The authority of interpretive communities* Cambridge, Ma.: Harvard University Press. Foucault, M. (1972). *The archaeology of knowledge*. London: Tavistock Publications.
Foucault, M. (1982a). Questions of method: an interview with Michel Foucault. *Ideology and Consciousness, 8(6),* 3-14.

Foucault, M. (1982b). The subject and the power. In H. Dreyfus & P. Rabinow (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics*, . Brighton: Harvester.
Foucault, M. (1988). *Politics, philosophy, culture: interviews and other writing*. New York: Routledge.
Foucault, M. (1992). *Discipline and punish*. London: Penguin.
Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18(9)*, 27-32.
Freud, S. (1963). *Civilisation and its discontents*. London: The Hogarth Press.
Friedenberg, E. (1969). , *Proceedings of the 1969 invitational conference on testing problems*, . Princeton: Educational Testing Service.
Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity, *Review of research in education*, (Vol. 20, pp. 337-391).
Garman, N. (1994). Qualitative enquiry: meaning and menace for educational researchers. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp.3-14). Adelaide: Flinders University of South Australia.
Garman, N., & Holland, P. (1995). the rhetoric of school reform reports: sacred, sceptical and cynical interpretations. In R. Ginsberg & D. Plank (Eds.), *Commissions, reports, reforms and educational policy*. Westport: Praeger.
Gillis, S., & Macpherson, C. (1995,). *Examination of the links between pre-employment qualifications and on the job competency based assessment.* Paper presented at the Australian Association for Research in Education, 25th Annual Conference, Hobart. Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.
Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement, 15(4)*, 237-261.
Golstein, H. (1979). Changing educational standards: A fruitless search. *Journal of the NAIEA*, 11(3), 18-19.
Gonzalez, E. J., & Beaton, A. E. (1994). The determination of cut scores for standards. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education*.
Good, F., & M, C. (1988). Grade awarding judgements in differential examinations. *British Educational Research Journal, 14(3)*, 263-281.
Green, M. (1994). Epistemology and Educational Research: the Influence of Recent approaches to Knowledge. *Review of Research in Education, 20*, 423-464.
Green, P. (1981). *The pursuit of inequality*. Oxford: Martin Robertson. Guba, E. (1990). *The paradigm dialog*. Newbury Park: SAGE Publications.
Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.
Hacking, I. (1991). How should we do the history of statistics? In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect*, . London: Harvester Wheatsheaf. Haertel E H. (1991). New forms of teacher assessment, *Review of Research in Education*, (Vol. 17, pp. 3-29).
Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.
Hambleton, R., & Zaal, J. (Eds.). (1991). *Advances in educational and psychological testing*. Boston: Kluwer Academic Publishers.
Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.
Hartog, P., & Rhodes, E. (1936). *The marks of examiners*. London: Macmillan and Co.
Harvey, L., & Greed, D. (1993). Defining quality. *Assessment and Evaluation in Higher Education, 18(1)*, 9-34.
Horkheimer, M., & Adorno, T. (1972). *Dialectic of enlightenment*. New York: Herder and Herder.

House, E. (1991). Evaluation and social justice. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp. 233-247). Chicago: University of Chicago Press.
Howe, K. R. (1994). Standards, assessment, and equality of educational opportunity.*Educational Researcher, 23(8),* 27-33.
Hulin, C., Drasgow, F., & Parsons, C. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illinios: Dow Jones-Irwin.
Huxley, A. (1950). *The perennial philosophy*. London: Chatto and Windus.
Illich, I. (1971). *Deschooling society*: Calder and Boyers Ltd.
Jackson, N. (1993). Competence: A game of smoke and mirrors? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training,* .Canberra: Australian College of Education.
Jaeger, R., & Tittle, C. (Eds.). (1980). *Minimum competency achievement testing*. Berkeley: McCutchen Publishing Corporation.
Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement, Third edition*. New York: American Council on Education, Macmillan Publishing Company.
Johnston, B., & Dowdy, S. (1988). *Teaching and assessing in a negotiated curriculum.* Melbourne: Robert Anderson and Ass.
Johnston, B., & Pope, A. (1988). *Principles and practice of student assessment*. Adelaide: South Australian Education Department.
Jones, L. (1971). The nature of measurement. In R. Thorndike (Ed.), *Educational measurement: second edition,* (pp. 335-355). Washington: American Council on Education.
Kavan, R. (1985). *Love and freedom*. London: Grafton Books.
Keeney, B. (1983). *Aesthetics of change*. New York: The Guilford Press.
Kennedy, K., Marland, P., & Sturman, A. (1995). *Implementing national curriculum statements and profiles: corporate federalism in retreat.* Paper presented at the Annual Conference of the Australian Association for Research in Education, Hobart, 26-30 November.
Knight, B. (1992). Theoretical and practical approaches to evaluating the reliability and dependability of national curriculum test outcomes, : Unpublished article.
Korzybski, A. (1933). *Science and sanity*. Lakeville, Con: International non-Aristotelian Pub. Co.
Laing, R. (1967). *The politics of experience*. Harmondsworth: Penguin.
Lather, P. (1991). *Getting Smart: Feminist research and pedagogy with/in the postmodern*. New York: Routledge.
Lazarus, M. (1981). *Goodbye to excellence: A critical look at minimum competency testing*. Boulder: Westview Press.
LeCompte, M., Millroy, W., & Preissle, J. (1992). *The handbook of qualitative research in education*. San Diego: Academic Press Inc.
Levin, H. (1978). Educational performance standards: Image or substance. *Journal of Educational Measurement, 15(4),* 309-319.
Lincoln, Y. (1990). The making of a constructivist. In E. Guba (Ed.), *The paradigm dialog*. Newbury Park: Sage Publications.
Lincoln, Y. (1995). Emerging criteria for quality in qualitative and interpretative research. *Qualitative Inquiry, 1*(3275-289).
Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23(9),* 4-14.
Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20(8),* 15-21.
Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Lorge, I. (1951). The fundamental nature of measurement. In E. Lindquist (Ed.), *Educational measurement,* (pp. 533-559). Washington: American Council on Education.

Madaus, G. F. (1986). Measurement specialists: Testing the faith - A reply to Mehrens. *Educational Measurement: Issues and Practice, 5(4),* 11-14.
Mager, R. (1962). *Preparing instructional objectives.* Palo Alto, CA: Feardon Publishers.
Marshall, C. (1990). Goodness criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.
Masson, J. (1991). *Final analysis*. London: Harper Collins.
Masters, G. (1994, 17 March). *Setting and measuring performance standards for student achievement.* Paper presented at the Public Investment in School Education: Costs and Outcomes, Canberra.
Maturana, H., & Guiloff, G. (1980). The quest for the intelligence of intelligence. *Journal of Social Biological Structures, 3.*
Mayer, C. C. (1992). *Putting general education to work: the key competencies report* Melbourne: Australian Educational Council and Ministers of Vocational Education, Employment and Training.
McDonald, R. (1994, October, 1994). Led astray by competence? Paper presented at the Australian National Training Authority, Brisbane.
McGovern, K. (1992). National competency standards - the role of the National Office of Overseas Skills Recognition. *The Journal of Teaching Practice, 12(1),* 33-46.
Meadmore, D. (1993). The production of individuality through examination. *British Journal of Sociology in Education, 14(1),* 59-73.
Meadmore, D. (1995). Linking goals of governmentality with policies of assessment. *Assessment in Education, 2(1),* 9-22.
Melton, R. (1994). Competencies in perspective. *Educational Research, 36(3),* 285-294.
Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.
Messick, S. (1989b). Meaning and values in test validation. *Educational Researcher, 18(2),* 5-11.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2),* 13-23.
Miller, A. (1983). *For your own good*. New York: Farrar, Straus, Giroux.
Miller, A. (1984). *Thou shalt not be aware*. London: Pluto Press.
Miller, C., & Parlett, M. (1974). *Up to the mark: a study of the examination game*. London: Society for Research into Higher Education.
Millman, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing company.
Mishler, E. (1986). *Research interviewing*. Cambridge: Harvard University Press.
Mitroff, I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of the social sciences (3),* 117-134.
Moss, P. A. (1992). Shifting concepts of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62(3),* 229-258.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher,23(2),* 5-12.
Mykhalovskiy, E. (1996). Reconsidering Table Talk: Critical thoughts on the relationship between sociology, autobiography and self-indulgence. *Qualitative Sociology, 19(1),* 131-151.
Nairn, A. (1980). *The reign of ETS* . Washington. National Training Board. (1992). *National Competency Standards: Policy and Guidelines (Second Edition)* . Canberra: National Training board.
National Training Board. (1995, January 1995). Who's doing what? Assessment in Australia today. *NTB Network - Special Conference Edition,* 19-20.
Norris, N. (1991). The trouble with competence. *Cambridge Journal of Education, 21(3),* 331-341.
Nuttall, D. (1979). The myth of comparability. *Journal of the NAIEA, 11(3),* 16-18.

Nuttall, D., Backhouse, J., & Willmott, A. (1974). *Comparability of standards between subjects.* (Vol. 29). Oxford: Evans/Methuen Educational.
Oakley, A. (1991). Interviewing women. In H. Roberts (Ed.), *Doing feminist research*, London: Routledge and Kegan Paul.
Orrell, J. (1996). Assessment in higher education: an examination of everyday academic's thinking-in-assessment, beliefs-about-assessment, and a comparison of assessment behaviours and beliefs. Unpublished Ph D, Flinders University of South Australia, Adelaide.
Partington, J. (1994). Double-marking students' work. *Assessment and Evaluation in Higher Education, 19(1),* 57-60.
Pawson, R. (1989). *A measure of measures*. London: Routledge.
Pearson, A. (1984). Competence: a normative analysis. In E. Short (Ed.), *Competence; Inquiries into its meaning and acquisition in educational settings,* (pp. 31-40). Lanham, MD: University Press of America.
Pennycuick, D., & Murphy, R. (1988). *The impact of graded tests*. London: The Falmer Press.
Perkins, D., & Salomon, G. (1988). Teaching for transfer. *Educational Leadership*(September), 22-32.
Persig, R. (1975). *Zen and the art of motorcycle maintenance: An enquiry into values*. New York: Bantam Press.
Persig, R. (1991). *Lila: An enquiry into morals*. London: Bantam Press.
Peters, M. (1996). *Poststructuralism, politics and education*. Westport: Bergin & Garvey.
Phillips, D. (1990). Subjectivity and objectivity: an objective enquiry. In E. Eisner & A.
Peshkin (Eds.), *Qualitative enquiry in education*. New York: Teachers college, Colombia University.
Popkewitz, T. (1984). *Paradigm and ideology in educational research*. London: The Falmer Press.
Porter, P., Rizvi, F., Knight, J., & Lingard, R. (1992). Competencies for a clever country: Building a house of cards? *Unicorn, 18(3),* 50-58. Prigogine, I., & Stengers, I. (1985). *Order out of chaos*. London: Fontana.
Quine, W. (1953). *From a logical point of view*. New York: Harper and Row. Rechter, B., & Wilson, N. (1968). Examining for university entrance in Australia: Current practices. *Quarterly Review of Australian Education, 2(2).*
Reilly, R., & Chao, G. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 33(1),* 1-55.
Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum and performance: A historical and comparative perspective. *Educational Researcher, 14(4),* 5-20.
Rorty, R. (1991). *Objectivity, relativism, and truth*. Cambridge: Cambridge University Press.
Rose, N. (1990). *Governing the soul: The shaping of the private self.* London: Routledge.
Rosenberg. (1967). *On quality in art: criteria of excellence, past and present*. Princeton: Princeton University Press.
Royal Commission. (1974). *Report on the suspension of a high school student* .Adelaide: South Australian Government.
Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13(2),* 191-209.
Sadler, R. (1995). Comparability of assessments, grades and qualifications. Paper presented at the AARE Conference, Hobart, 24 November.
Schmidt, F., Hunter, J., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: a red herring. *Journal of Applied Psychology, 66(2)*166-185.
Schnell, J. (1980). *The fate of the earth*. London: Picador.
Schwandt, T. (1990). Paths to enquiry in the social disciplines. In E. Guba (Ed.), *The paradigm dialog,* . Newbury Park: SAGE Publications.
Scriven, M. (1991). *Evaluation thesaurus; fourth edition*. Newbury Park, Cal: SAGE Publications.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20(7)*, 2-16.
Shepard, L. A. (1993). Evaluating test validity, *Review of research in education, 19*.
Sherman, R., & Webb, R. (1988). *Qualitative research in education*. New York: Falmer.
Slater, P. (1966). *Microcosm*. New York: John Wiley.
Smith, B. (1994). Addressing the delusion of relevance: Struggles in connecting educational research and social justice. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp. 43-56). Adelaide: Flinders University of South Australia.
Smith, J. (1990). Alternative research paradigms and the problem of criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.
Smith, J. (1993). *After the demise of empiricism: the problem of judging social and education inquiry*. Norwood, N.J.: Ablex Publishing Corporation.
Smyth, J. (Ed.). (1994). *Qualitative approaches in educational research*. Adelaide: Flinders University of South Australia.
Soucek, V. (1993). Is there a need to redress the balance between systems goals and life world-oriented goals in public education in Australia? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training*. Canberra: Australian college of Education.
Spearritt, D. (Ed.). (1980). *The improvement of measurement in education and psychology*. Hawthorne, Victoria: Australian Council for Educational Research. Stake, R. (1991). The countenance of educational evaluation. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education: at quarter century*, (pp. 67-88). Chicago: the University of Chicago Press.
Stanley, G. (1993). The psychology of competency-based education. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.
Stern, D. (1991). *Diary of a baby*. London: Fontana.Sternberg, R. (1990). T & T is an explosive combination: technology and testing. *Educational Psychologist, 25(3&4)*, 201-222.
Sydenham, P. (1979). *Measuring instruments: tools of knowledge and control*. London: Peter Peregrinus Ltd.
Taylor, C. (1994). Assessment for measurement of standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31(2)*, 231-262.
Taylor, P. (1961). *Normative discourse*. Englewood Cliffs: Prentice-Hall, Inc. The Flinders University of South Australia. (1997). *Calender*. Adelaide: Flinders University of South Australia.
The Social Development Group. (1979). *Developing the classroom group: How to make your class a better place to live in* . Adelaide: South Australian Education Department.
The Social Development Group. (1980). *How to make your classroom a better place to live in* . Adelaide: South Australian Education Department.
Thompson, P., & Pearce, P. (1990). *Testing times*. Adelaide: TAFE National Centre for Research and Development Thompson, W. (Ed.). (1987). *Gaia, a way of knowing*. Hudson: Lindisfarne Press.
Travers, E., & Allen, R. (1994). *Random sampling of student folios: a pilot study (10)*. Brisbane: Board of Senior Secondary School Studies, Queensland. Watzlewich, P. (1974). *Change*. New York: W Norton & Co.
Weiss, C. (1991). Evaluation research in the political context. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education; at quarter century*, (pp. 211-231). Chicago: The University of Chicago Press.
Wheeler, L. (1993). Reform of Australian vocational education and training: A competency-based system. In c. Collins (Ed.), *Competencies: the competencies debate education and training*. Canberra: Australian College of Education.
Wiggins, G. (1988). Teaching to the (authentic) test. *Educational Leadership*(September), 41-47.
Wilbur, K. (1977). *The spectrum of consciousness*.

Wheaton: Quest.Wilbur, K. (1982). *Up from Eden: A transpersonal view of human evolution*. Boston:Shambhala.
Wilbur, K. (1991). *Grace and grit*. North Blackburn: Collins Dove.
Wilbur, K. (1995). *Sex, ecology, spirituality*. Boston: Shambhala.
Wiliam, D. (1995). Technical issues in criterion-referenced assessment: evidential and consequential bases. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.
Williams, F. (Ed.). (1967). *Educational evaluation as feedback and guide*. Chigago: The National Society for the Study of Education.
Willmott, A. S., & Nuttall, D. L. (1975). *The reliability of examinations at 16+*. London: Macmillan Education Ltd.
Wilson, N. (1966). *A programmed course in physics, Form V*. Sydney: Angus and Robertson.
Wilson, N. (1969). Group discourse and test improvement. Unpublished data.
Wilson, N. (1969). A study of test-retest and of marker reliabilities of the 1966 commonwealth secondary scholarship examination. *ACER Information Bulletin, 50(1).*
Wilson, N. (1970). *Objective tests and mathematical learning.* Melbourne: Australian Council for Educational Research.
Wilson, N. (1972). *Assessment in the primary school*. Adelaide: South Australian Education Department.
Wilson, N. (1974). *A framework for assessment in the secondary school* . Adelaide:South Australian Education Department.
Wilson, N. (1985). *Young people's views of our world* (Peace Dossier 13). Melbourne: Victorian Association of Peace Studies.
Wilson, N. (1986). Programs to reduce violence in schools. Adelaide: South Australian Education Department.
Wilson, N. (1992). *With the best of intentions*. Nairne: Noel Wilson.
Withers, G. (1995). Achieving comparability of school-based assessments in admissions procedures to higher education. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.
Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment, *Review of research in education*, (Vol.17, pp. 31-71).
Wolf, R. M. (1994). The validity and reliability of outcome measures. In A. C. Tuijnman & T. Neville Postlethwaite (Eds.), *Monitoring the standards of education.*
Wood, R. (1987). Aspects of the competence-performance distinction: Educational, psychological and measurement issues. *Curriculum Studies, 19(5),* 409-424.
Wood, R. (1987). *Measurement and assessment in education and psychology*. London: The Falmer Press.

Annexure 1.

Survey Questionnaire for Students:

(NOTE): The below information is solely gathered for academic purposes to conduct a research on Examination & Evaluations systems in Pakistan, this survey is specifically conducted to complete my M.Phil Thesis and in general to acquire information about the academic standards in Pakistan.

Instructions for participants:

Read and tick the information carefully, select a single option stated in bold, do not overwrite or cross the options.

Previous medium of Instruction	English	Urdu		
Acquired Education level	FA/Fsc	A-levels		
Age group	15---------18	19----------23		
Board of Intermediate	Punjab Board	Federal Board	Sindh Board	FATA
Marks secured in NTS:	50% or below	60% or above	80% or above	
Cleared NTS Test (NAT 1) in:	1st attempt	2nd attempt	3rd attempt	
Current field of study:	Engineering	Management Sciences	Others	

(Annexure 1.) Questionnaire sample for Summative/Formative Evaluation (Pupils feedback to Identify Non-standardization in local assessment tests in Pakistan (NAT-1)

Instructions:
Please read each statement carefully and answer according to the following options:

1= Strongly Disagree
2= Disagree
3= Neutral
4= Agree
5= Strongly Agree

1).Do you think that NTS Tests have really contributed to improve your English language skill?	SD	D	N	A	SA
2).Do you think that the NTS Tests are merely a formality & do not focus on your university/college English language needs?					
3).Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?					
4).If you disagree with Q.no.3, do you think that NTS is a reliable & transparent assessment system?					
5).Are you satisfied with the assessment score awarded to you by NTS?					
6).If not satisfied because you think transcript was not properly checked?					
7).If not satisfied because you think English language assessment was too difficult?					

8) .you are not satisfied because you think that your numerical ability is stronger than your language ability?					
9).Do you think that the English language multiple choice questions are sufficient to assess your English language proficiency?					
10).Do you think that the English language test pattern of NTS is not focusing on your creative writing ability?					
11).Do you think that the English language test pattern of NTS is not sufficient to prepare you for any International language proficiency tests like: IELTS, TOEFL,GMAT or GRE?					
12). Do you think that the English language courses in your university curriculum really fulfill your academic needs?					
13). Do you think that the NTS entry tests are really difficult to acquire admission in universities?					
14).Do you think that every university must conduct it's own English language entry test?					
15). Are you satisfied with the grade you get in your English language assessments in class?					
16). If you do not get good grades in the English language tests, is it because of your poor previous academic background?					
17). Do you think that you are not getting good grades, is because of the boring curriculum scheme?					
18). Do you think that you are not getting good grades, is because you lack the skills required for attempting questions?					
19). Do you think that you are not getting good grades, is because you find the multiple choice questions difficult?					
20). Do you think that you are not getting good grades, is due to infrequent practice?					
21). Do you think that you are not getting good grades, is due to ineffective reading & writing skills?					
22). Do you think that the NTS entry test language portion is inadequate because it does not assess listening and speaking skills?					
23). Do you think that there should be a separate local English language test, as university entry criteria?					
24).Do you think that the local English language test would be beneficial for the Pakistani students in future?					

Annexure 2

Survey Questionnaire for Teachers:

(NOTE): The below information is solely gathered for academic purposes to conduct a research on Examination & Evaluations systems in Pakistan, this survey is specifically conducted to complete my M.Phil Thesis and in general to acquire information about the academic standards in Pakistan.
Kindly state your brief bio-data accordingly the below asked queries:
English language Teaching experience: (5---10 years) (10—15 years) (15—20 years).

Academic Qualification: M.A/Mhil. English Literature, M.A/ MPhil. Linguistics, M.A/MPhil. ELT, TESOL,CELTA, DELTA.

Acquired any pre-service training in Testing & Evaluation: (YES) (NO).

Acquired any In-service training in Testing & Evaluation: (YES) (NO).

State any institutional or other training session you attended :

Have you ever get a chance to become an external examiner on the behalf of National Testing systems (NTS) Pakistan: (YES) (NO).

Teaching English language at: (University) (College)

Name English language courses you are teaching currently:

1.----------
2.----------
3.----------
4.----------
5.----------
6.----------

Instructions:

Please read each statement carefully and answer according to the following options:

1. Strongly Disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly Agree

	SD	D	N	A	SA
1. Do you think that the NTS Tests have really contributed to improve your students English language skill?					
2. Do you think that the NTS Tests are merely a formality & do not focus on their university/college English language need?					
3. Do you think that the Universities/colleges must not be given an authority to take their aptitude tests?					
4. If you disagree with question no.3, NTS is a reliable & transparent assessment system?					
5. Do you think that your students justify the score they acquired in NTS?					

6. you are not satisfied by students' class performance because you think they are not properly assessed by NTS?					
7. you are not satisfied because you think that the English language assessment was too difficult in NTS?					
8. you are not satisfied because you think that students' numerical ability is stronger than their language ability?					
9. Do you think that the English language multiple choice questions are sufficient to assess students' English language proficiency?					
10. Do you think that the English language test pattern of NTS is not focusing on students' creative writing ability?					
11. Do you think that the English language test pattern of NTS is not sufficient to prepare students for any International language proficiency tests like: IELTS, TOEFL,GMAT or GRE?					
12. Do you think that the English language courses in your university curriculum really fulfill student's academic needs?					
13. Do you think that the NTS entry tests are really difficult for students to acquire admission in universities?					
14. Do you think that every university must conduct it's own English language entry test?					
15. Are you satisfied with the grades your students get in your English language assessments in class?					
16. If the students do not get good grades in the English language tests, is because of their poor previous academic background?					
17. The students are not getting good grades because of the boring curriculum scheme.					
18. The Students are not getting good grades because they lack the skills required for attempting questions.					
19. The students are not getting good grades because they find multiple choice questions difficult.					
20. The students are not getting good grades due to infrequent practice.					
21. The students are not getting good grades is due to 22. Ineffective reading & writing skills?					
23. Do you think that the NTS entry test language portion is inadequate because it does not asses listening and speaking skills test?					
24. Do you think that there should be a separate local English language test as university entry criteria?					
25. Do you think that the local English language test would be beneficial for Pakistani students in future?					

Annexure:3

5. Even if you do not _____ what I have to say, I will appreciate you listening to me with an open mind.
   (A). Concur with (B). Anticipate
   (C). Reject (D). Clarify

□ **Each of the following analogy question presents a related pair of words linked by a colon. Select the pair of words whose relationship is most like the relationship expressed in original pair.**

6. STARE : GLANCE
   (A). gulp : sip (B). scorn : admire
   (C). hunt : stalk (D). participate : observe

7. FIRE : ASHES
   (A). water : waves (B). event : memories
   (C). regret : melancholy (D). wood : splinters

8. SIGN : ZODAIC
   (A). poster : billboard (B). letter : alphabet
   (C). prediction : prophecy (D). signal : beacon

□ **Select a word or phrase from the options which is the most nearly opposite in meaning to the given capitalized word.**

9. COMPLY
   (A). Simplify (B). Strive
   (C). Rebel (D). Unite

10. AUTONOMY
   (A). Dependence (B). Animation
   (C). Renown (D). Altruism

11. TAPER
   (A). emphasize (B). broaden
   (C). split (D). restore

12. HALLOW
   (A). keep silence (B). prove incorrect
   (C). desecrate (D). accuse openly

□ **Read the passage below carefully and answer the questions on the basis of what is stated or implied.**

It is wonderful to observe how a baby grows. In its infancy, it responds only to pain and hunger. By the second month it can raise its head to look at things around it. It also begins to smile at people. By the time the baby is four months old, it begins to catch at things. By the time the baby is five months old, it can grasp objects and can even feed itself. At about seven months it can crawl. In late infancy its interests increase. It begins to like games, songs and even books.

13. In late infancy the baby takes interest in
(A). playing games (B). reading books
(C). different activities (D). singing songs

14. In what manner do you think the baby responds to hunger?
(A). shouting (B). laughing
(C). crying (D). shrieking

15. At five months of age, the baby can feed itself with
(A). considerable help from the mother
(B). little help from mother
(C). frequent help from mother
(D). a little help from mother

16. "By seventh month baby can crawl" means
(A). the baby can run
(B). the baby can walk on two legs
(C). the baby can run on all fours limbs
(D). the baby can walk on all four limbs

17. In the passage 'catch at' means to
(A). Hold (B). Get
(C). Try to grasp (D). Jump at

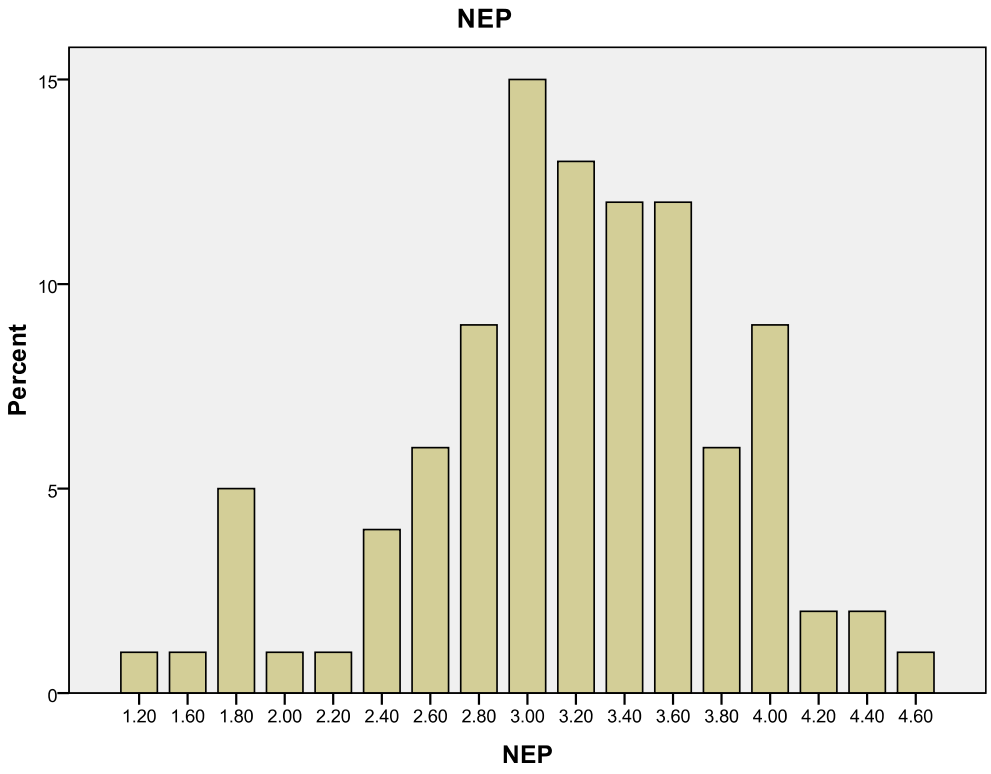
□ **Select a word or phrase from the options which is the most nearly similar in meaning to the given capitalized word.**

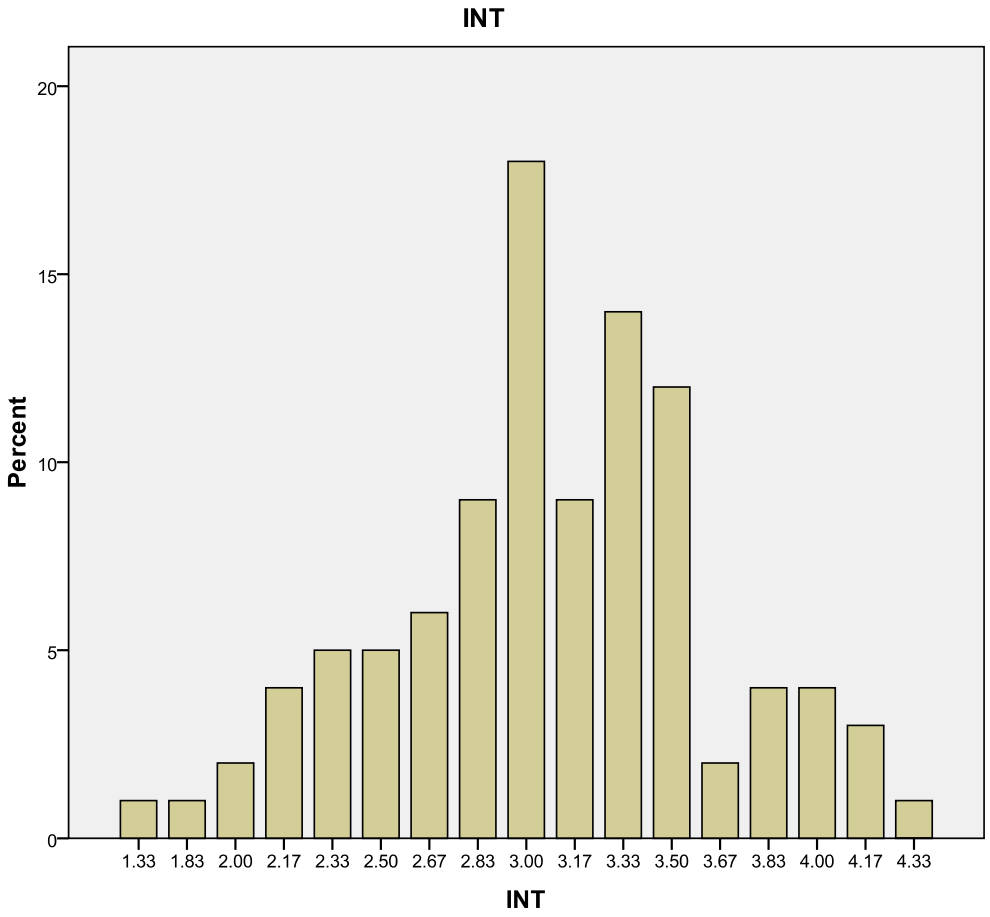
18. CONCEIT
(A). Pride (B). Wrath
(C). Humility (D). Inspiration

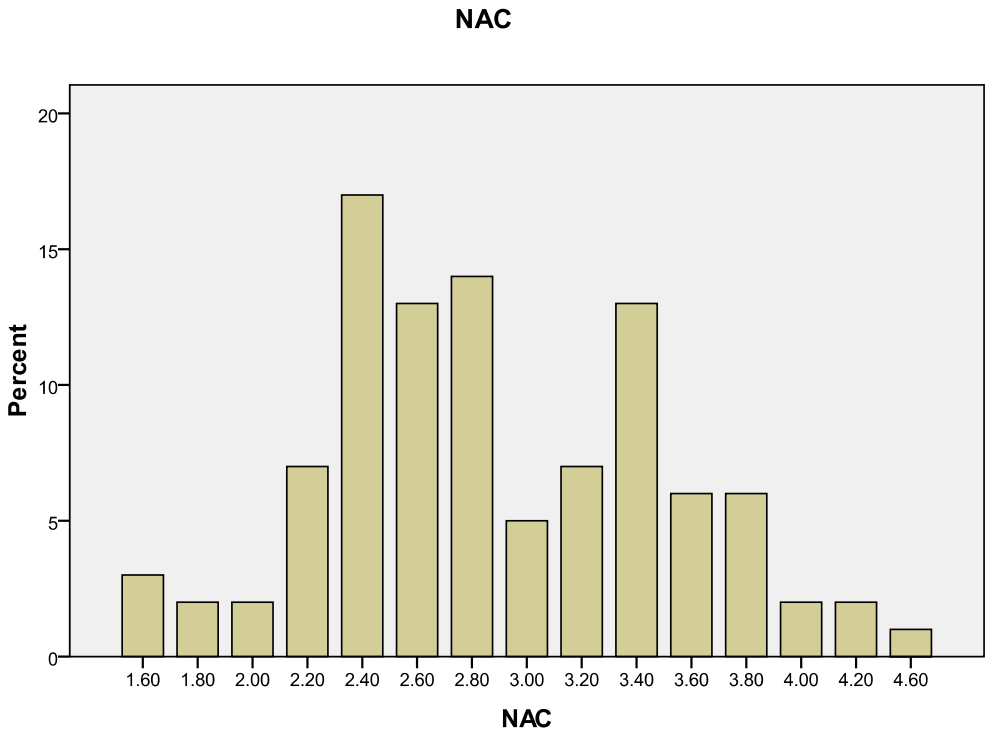
19. FIDELITY
(A). Friendly (B). Intrigue
(C). Loyalty (D). Faithlessness

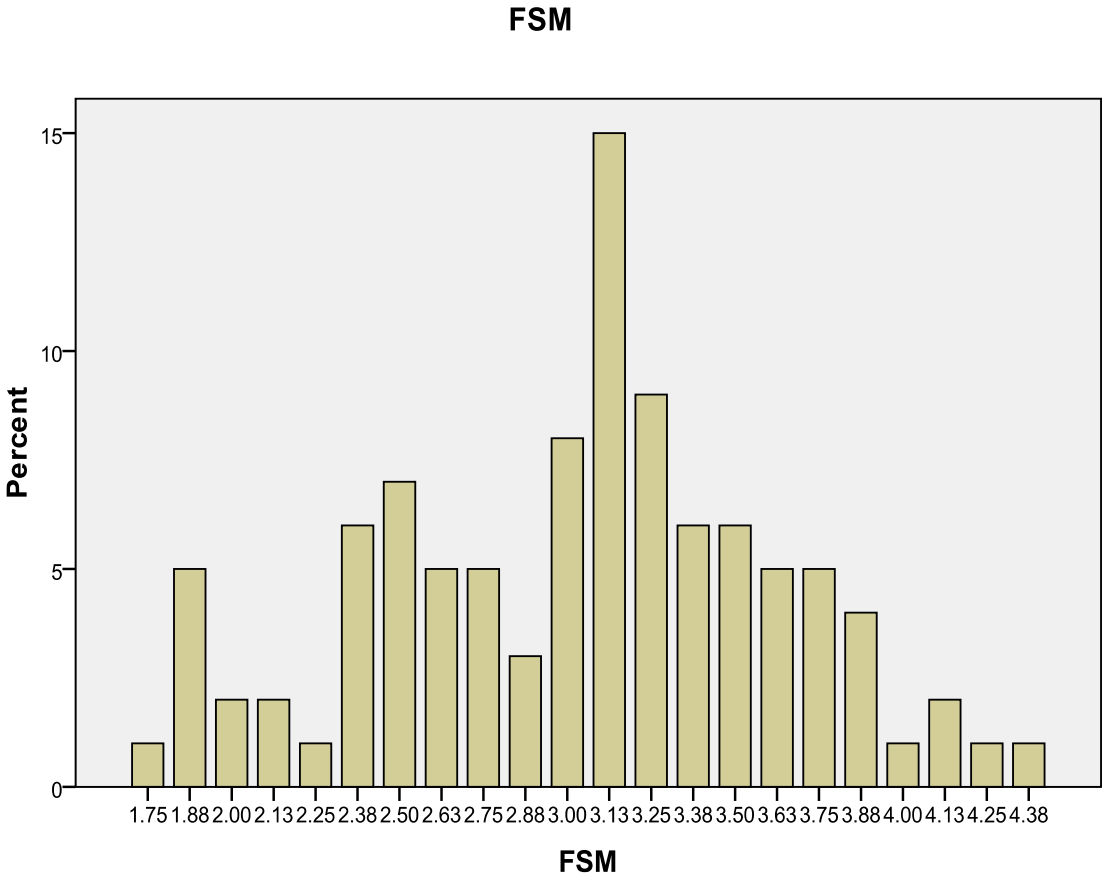
Annexure: 4

Students survey results in Bar Charts.



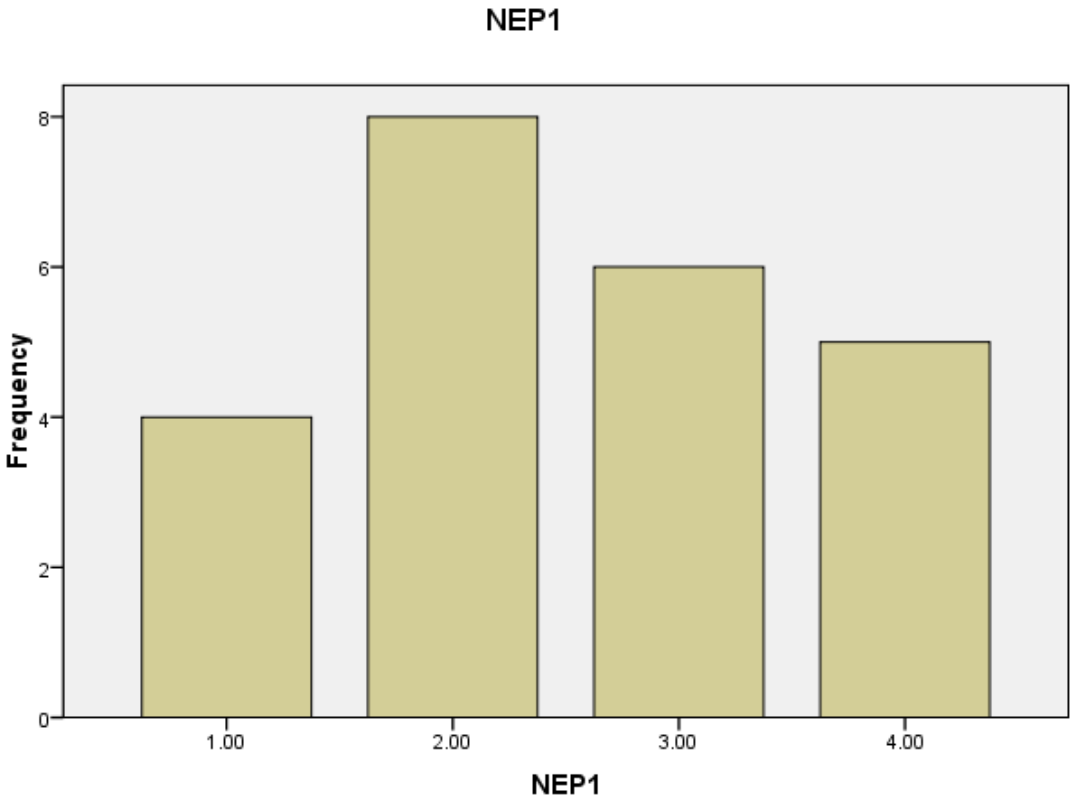

INT
Percent
20
15
10
5
0
1.33
1.83
2.00
2.17
2.33
2.50
2.67
2.83
3.00
3.17
3.33
3.50
3.67
3.83
4.00
4.17
4.33
INT

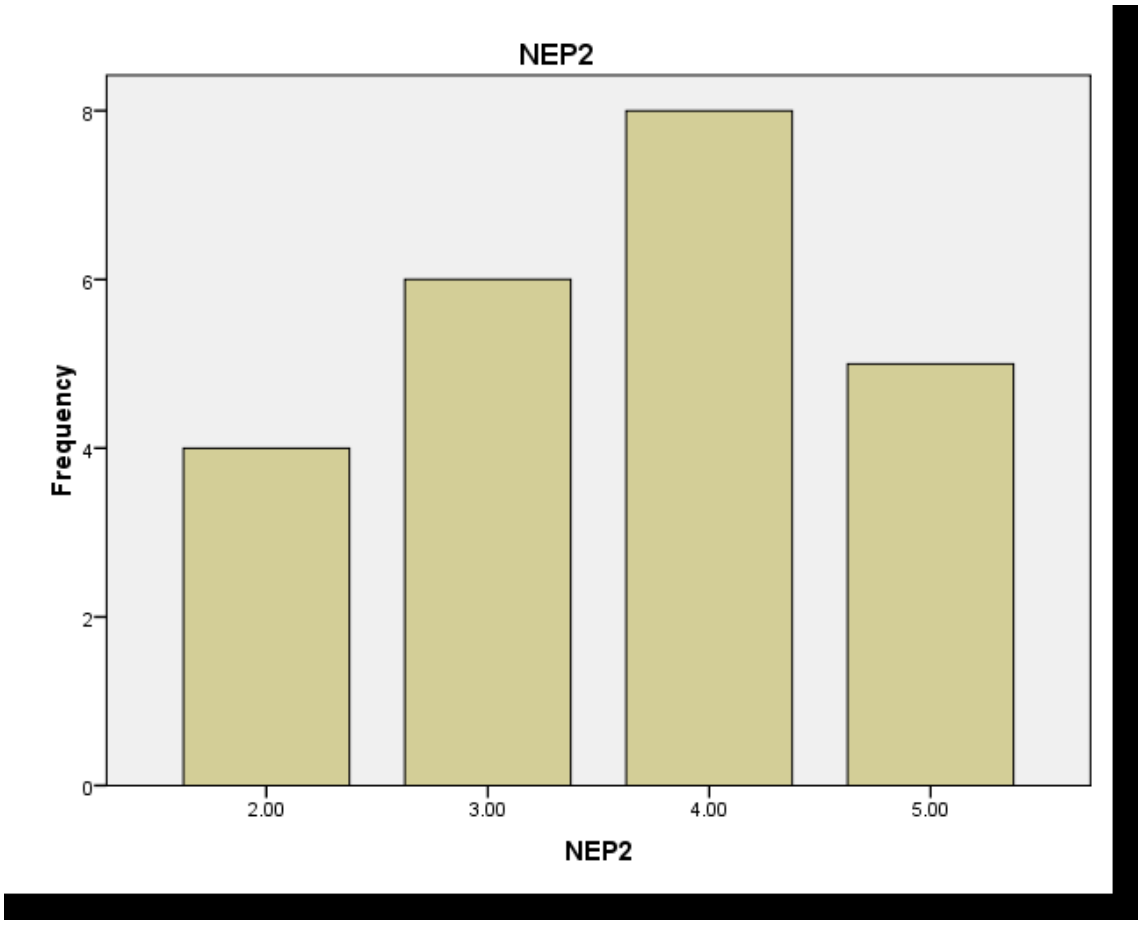

NAC
Percent
20
15
10
5
0
1.60
1.80
2.00
2.20
2.40
2.60
2.80
3.00
3.20
3.40
3.60
3.80
4.00
4.20
4.60
NAC

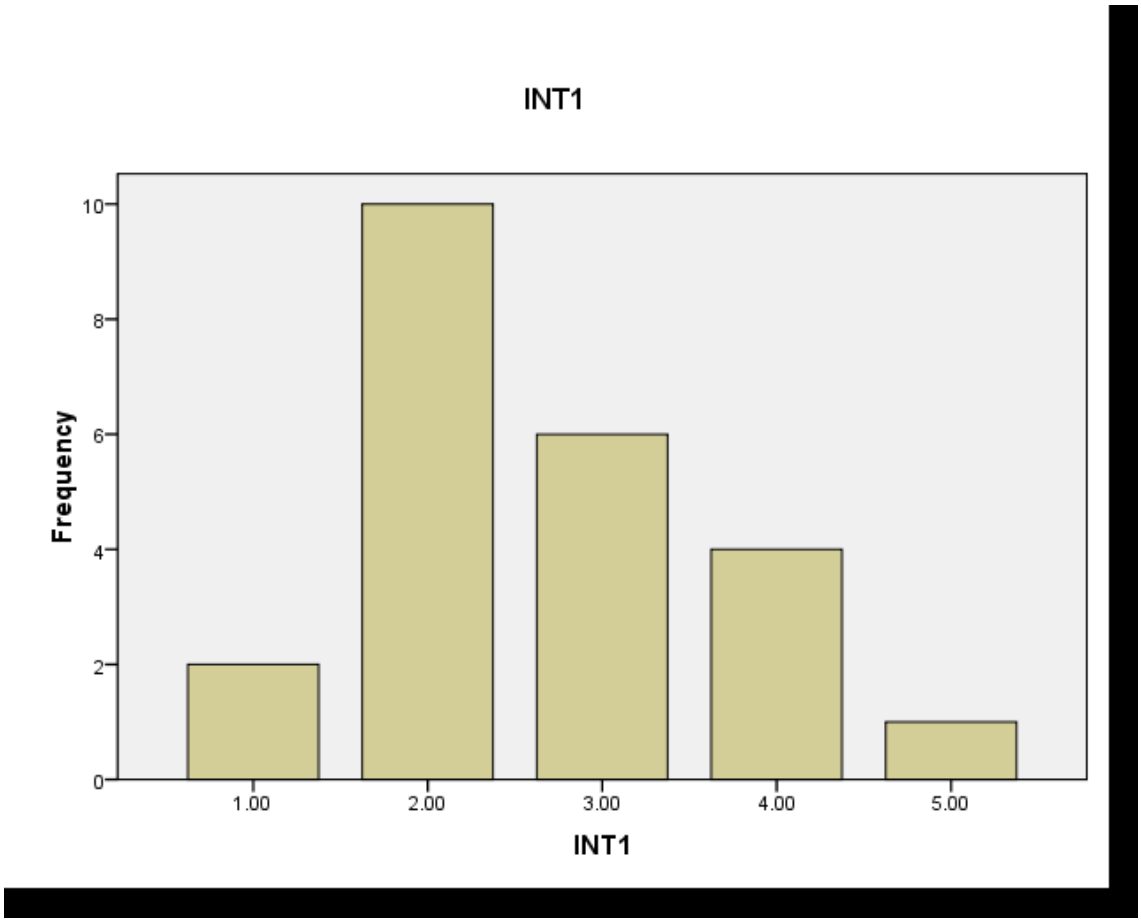

FSM
Percent
15
10
5
0
1.75
1.88
2.00
2.13
2.25
2.38
2.50
2.63
2.75
2.88
3.00
3.13
3.25
3.38
3.50
3.63
3.75
3.88
4.00
4.13
4.25
4.38
FSM

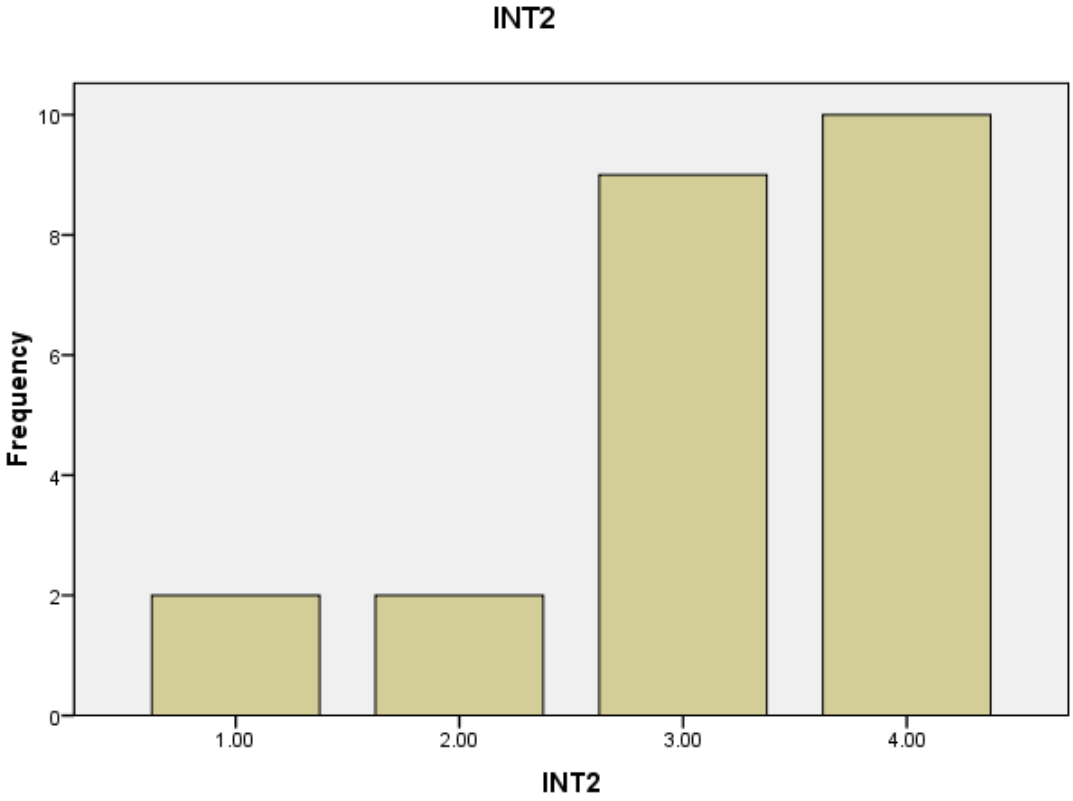
Annexure: 5

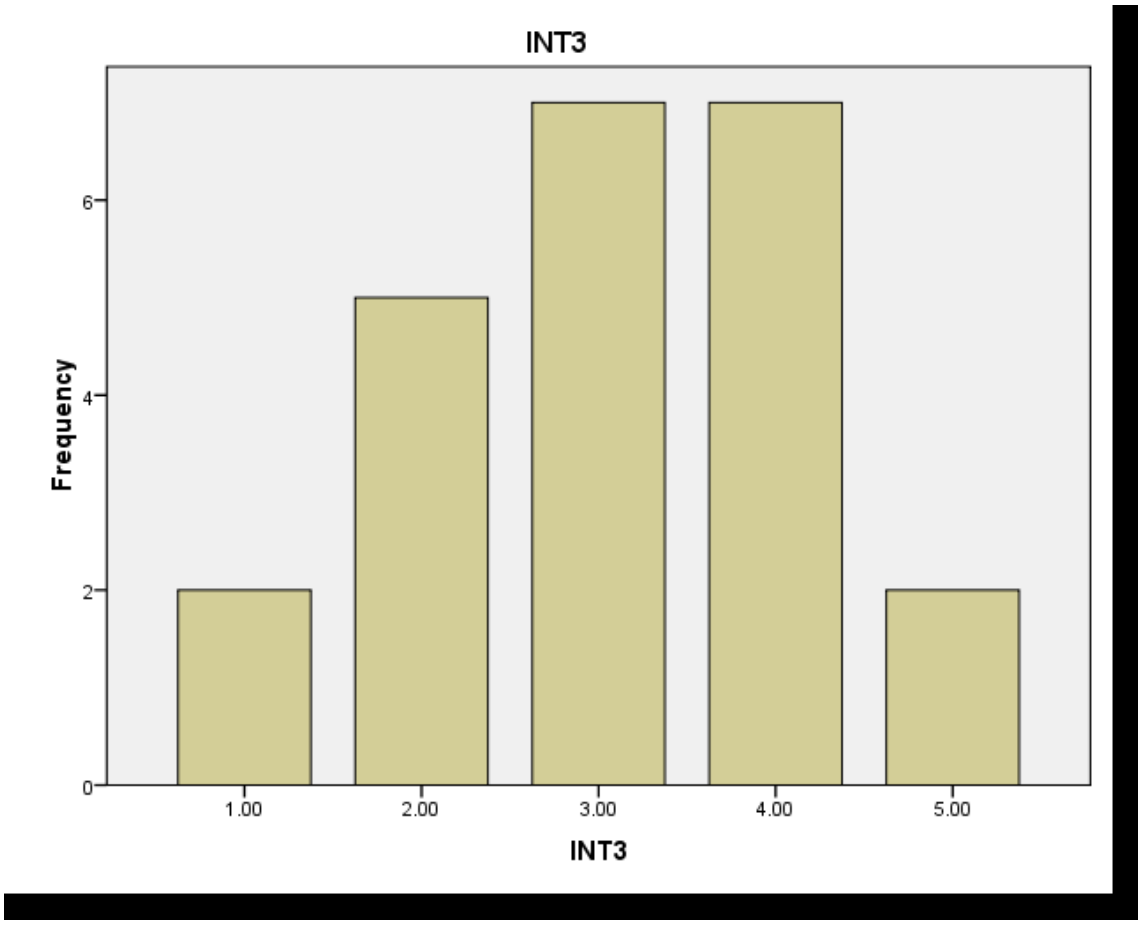
Teachers survey results in Bar Charts:

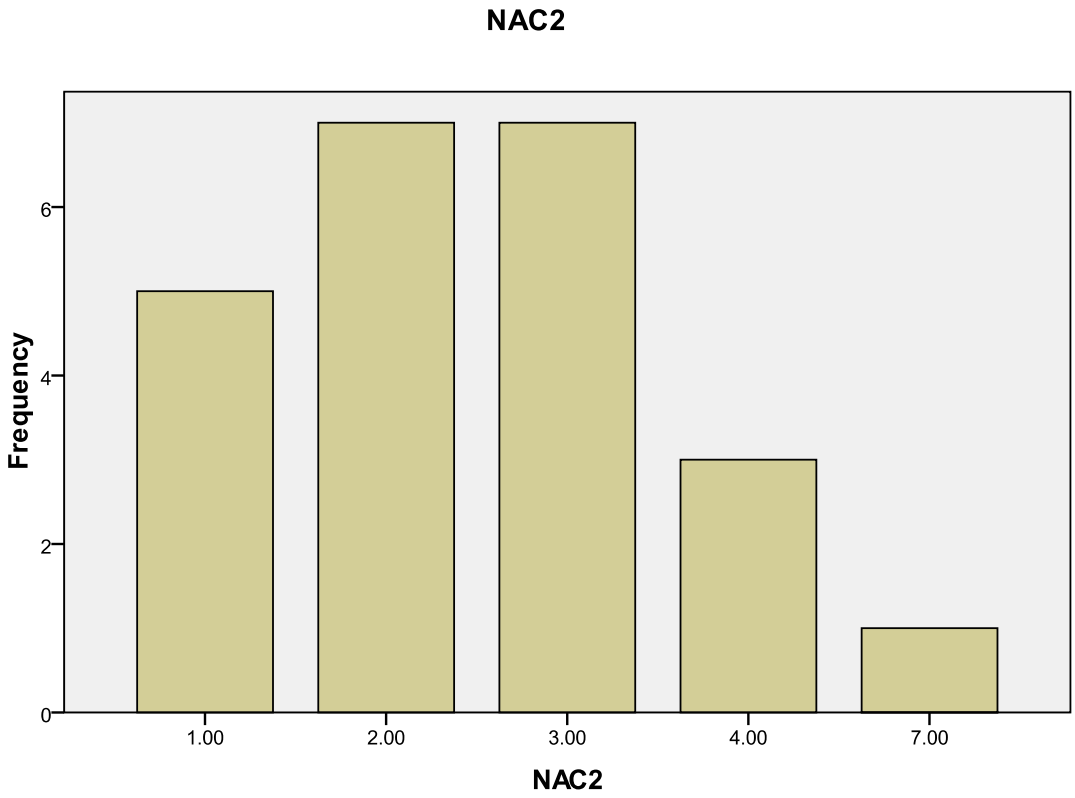


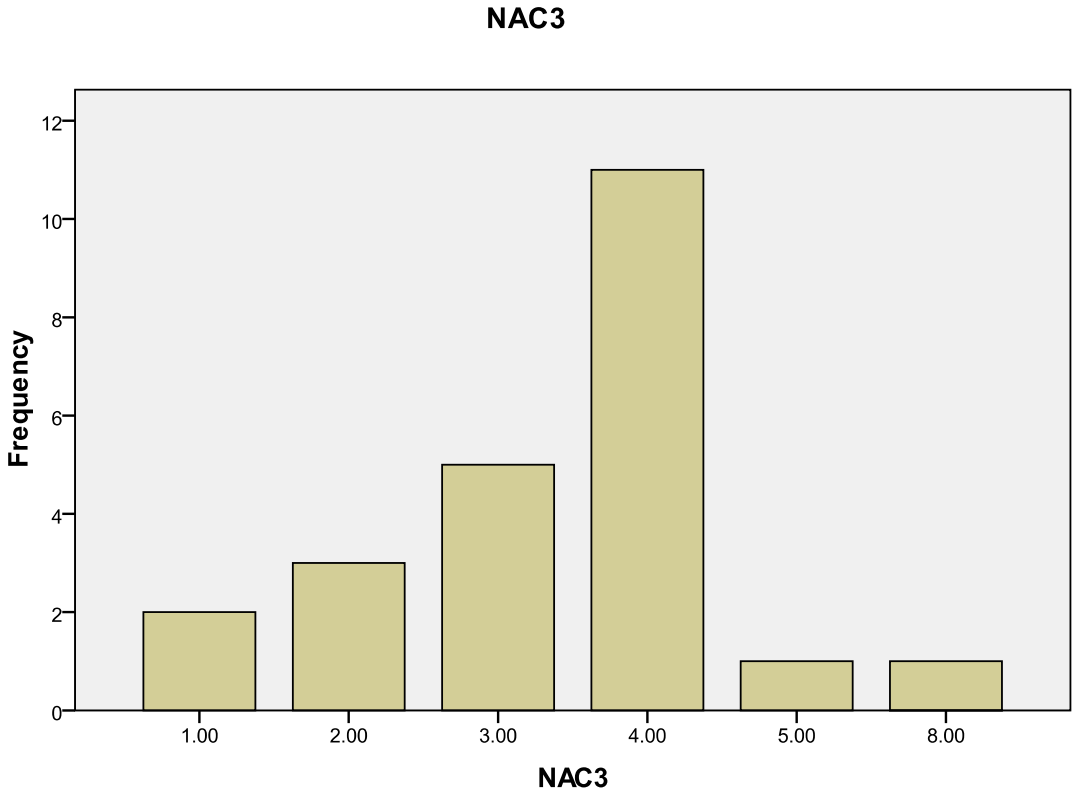

NEP2
Frequency
8
6
4
2
0
2.00
3.00
4.00
5.00
NEP2

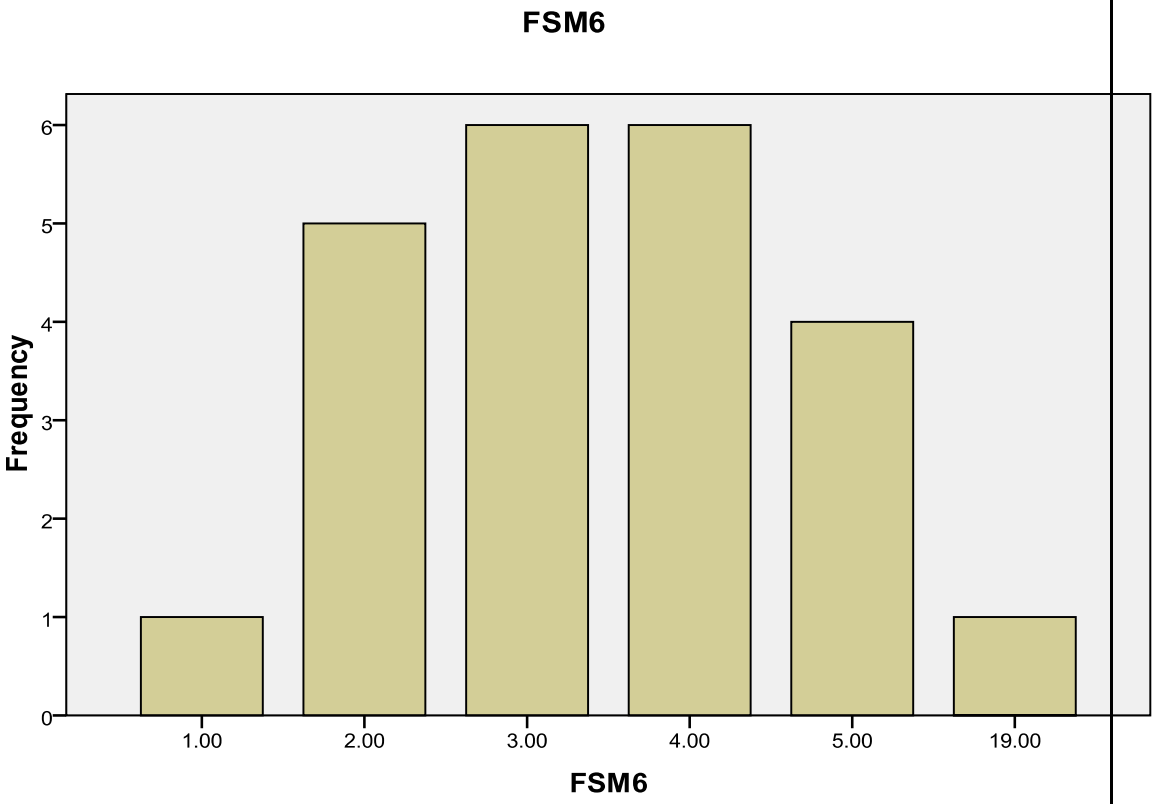

INT1
Frequency
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
INT1

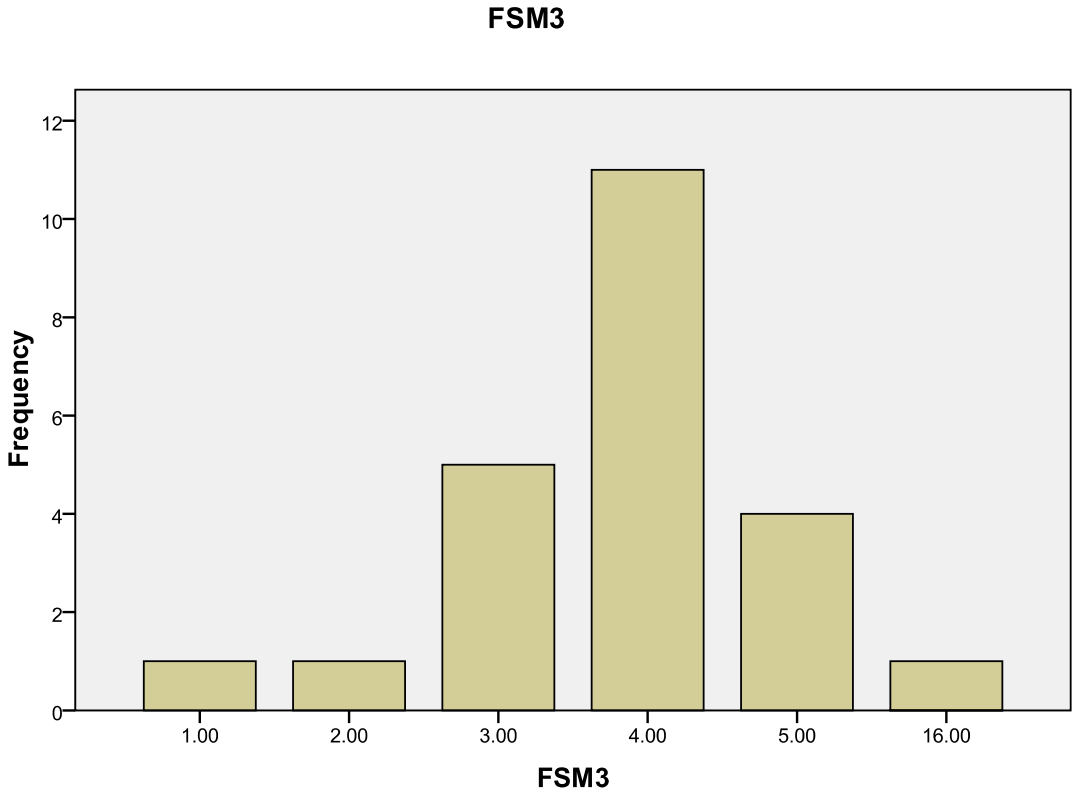

INT2
Frequency
10
8
6
4
2
0
1.00
2.00
3.00
4.00
INT2

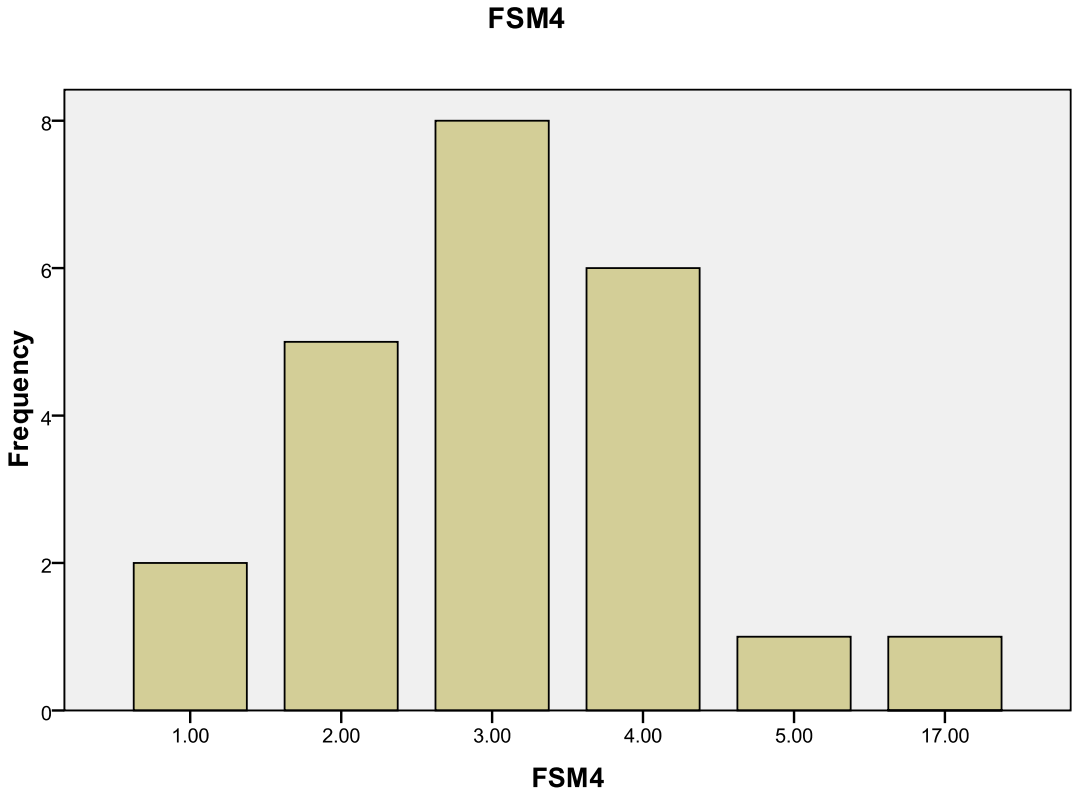

INT3
Frequency
6
4
2
0
1.00
2.00
3.00
4.00
5.00
INT3

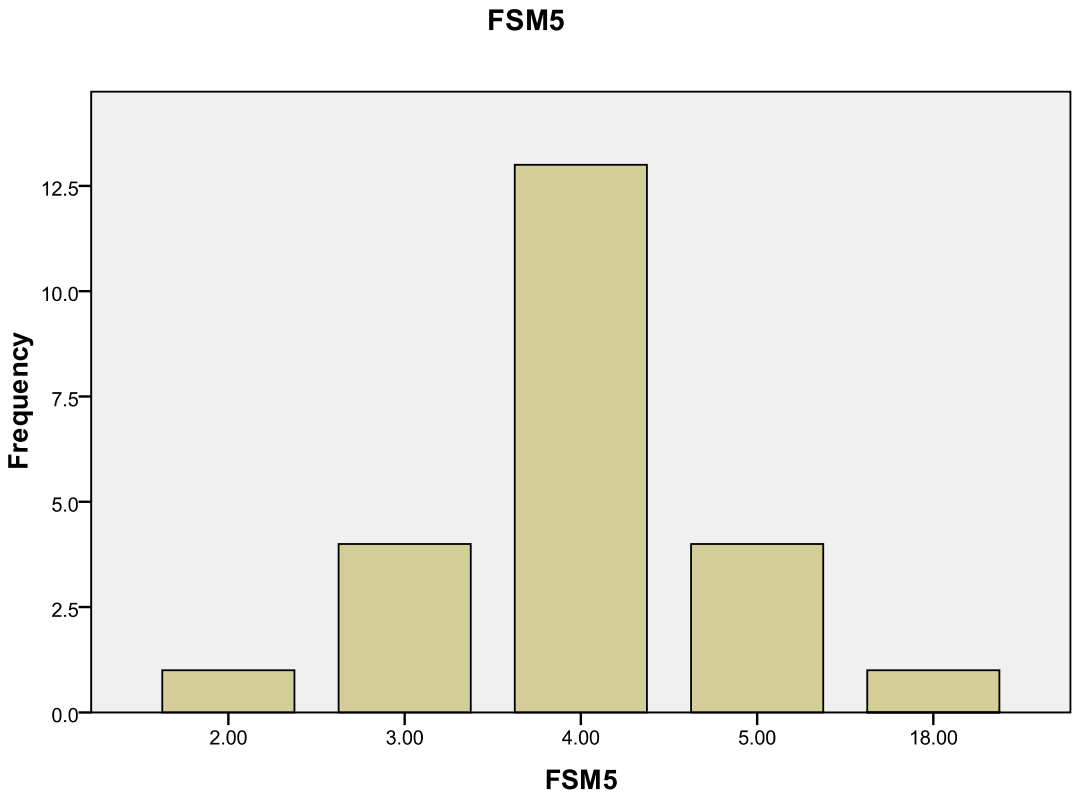

NAC2
Frequency
6
4
2
0
1.00
2.00
3.00
4.00
7.00
NAC2

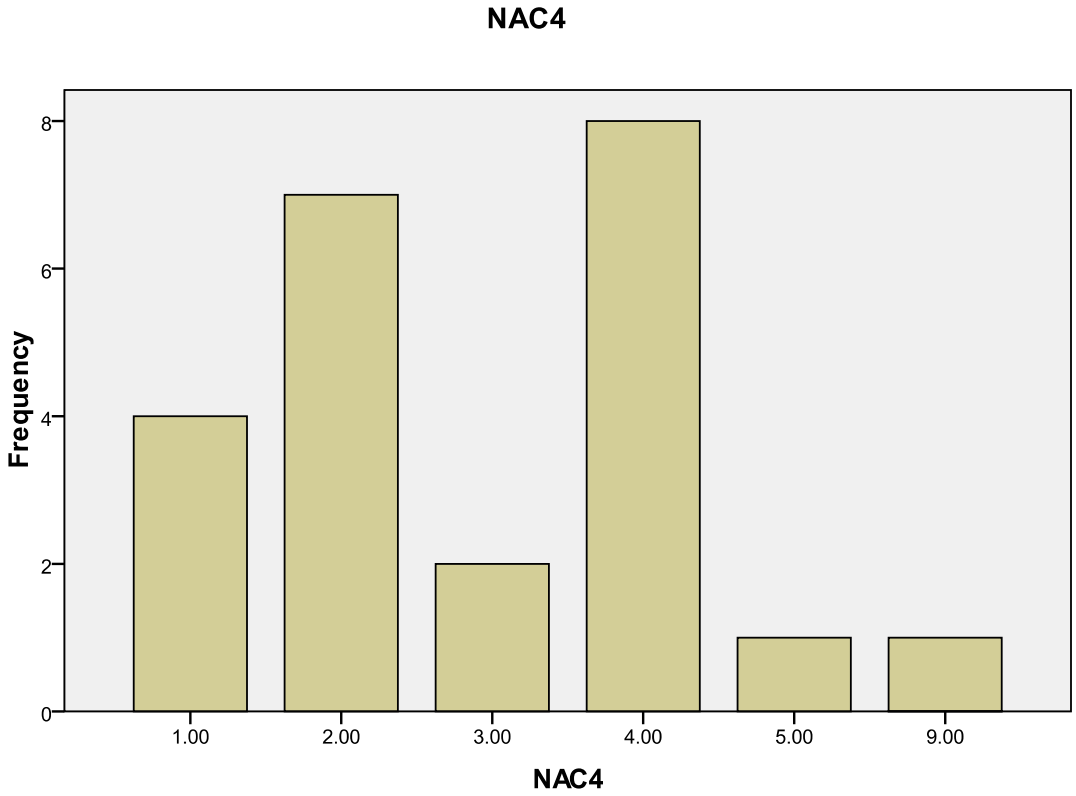

NAC3
Frequency
12
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
8.00
NAC3

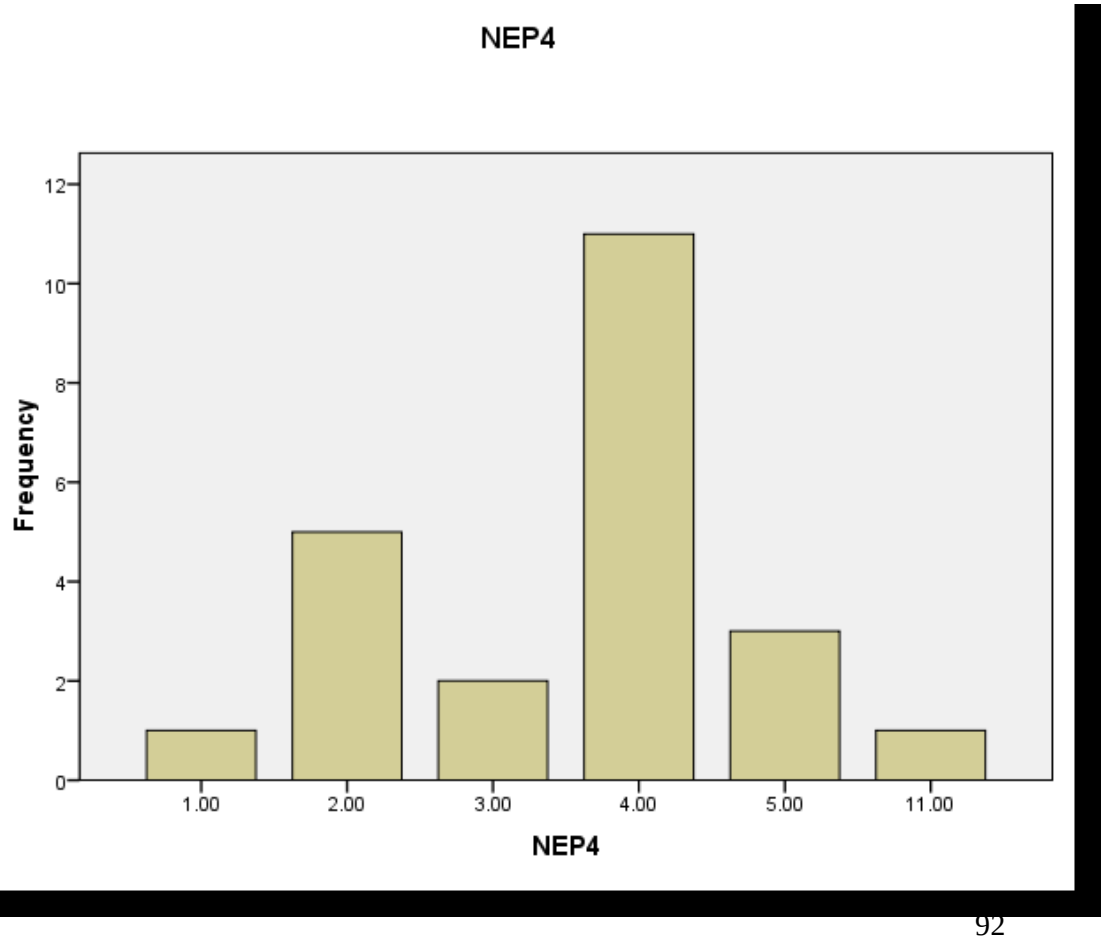

FSM6
Frequency
0
1
2
3
4
5
6
1.00
2.00
3.00
4.00
5.00
19.00
FSM6

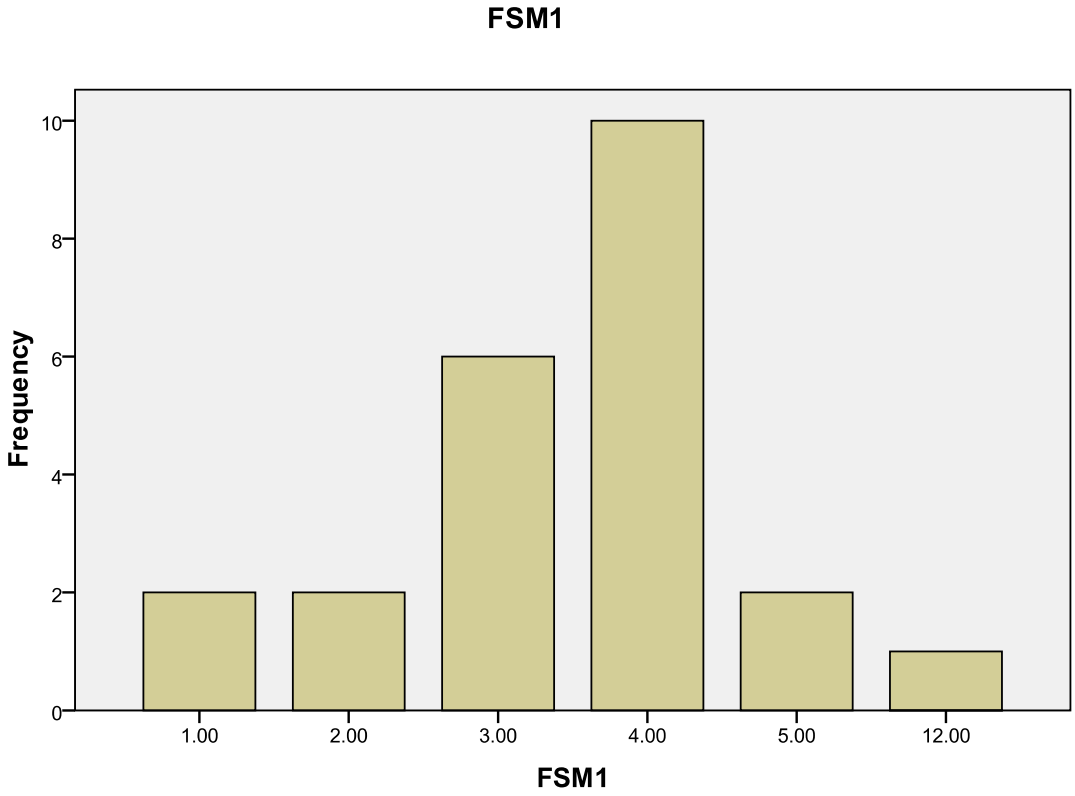

FSM3
Frequency
12
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
16.00
FSM3

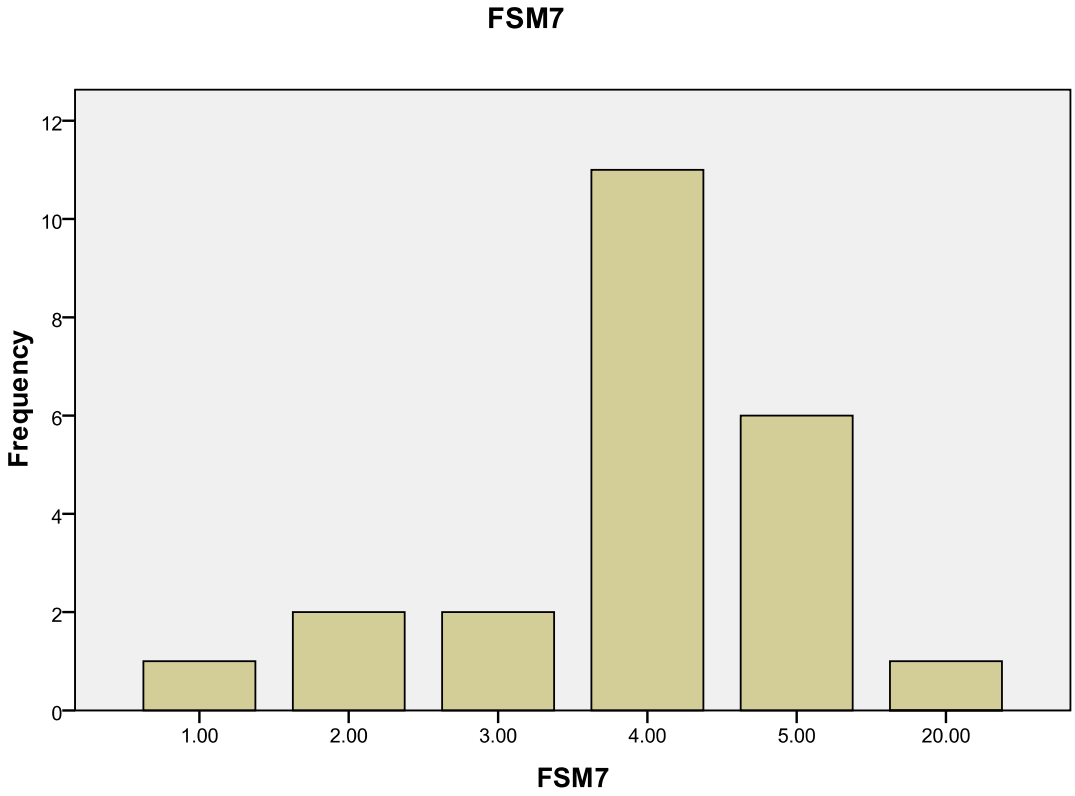

FSM4
Frequency
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
17.00
FSM4

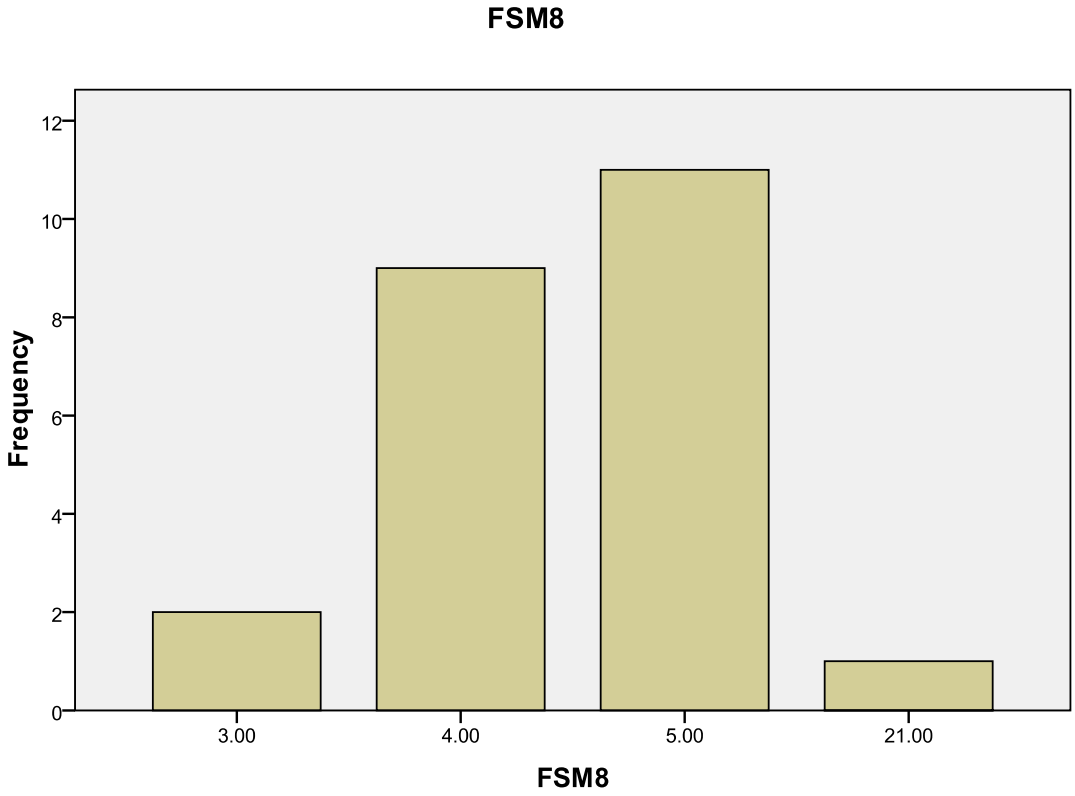

FSM5
Frequency
12.5
10.0
7.5
5.0
2.5
0.0
2.00
3.00
4.00
5.00
18.00
FSM5

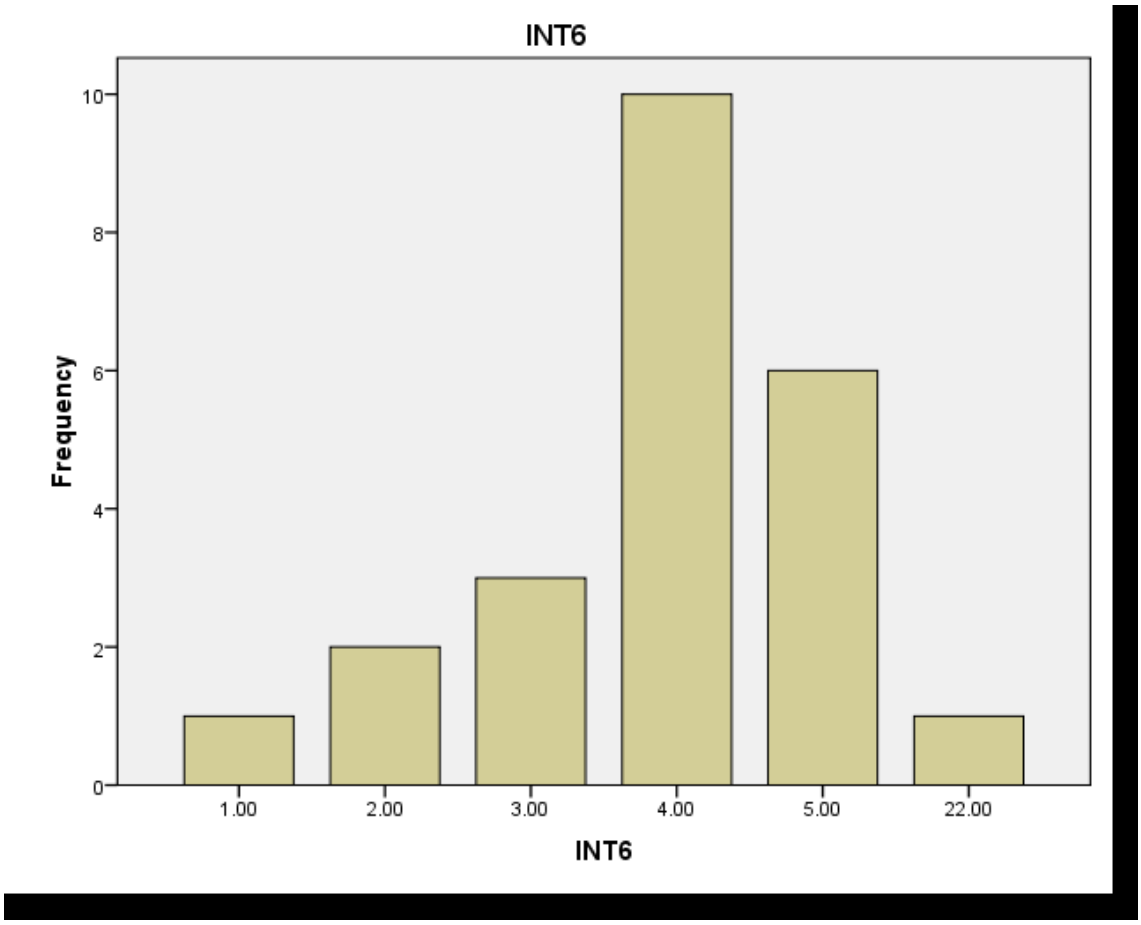

NAC4
Frequency
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
9.00
NAC4

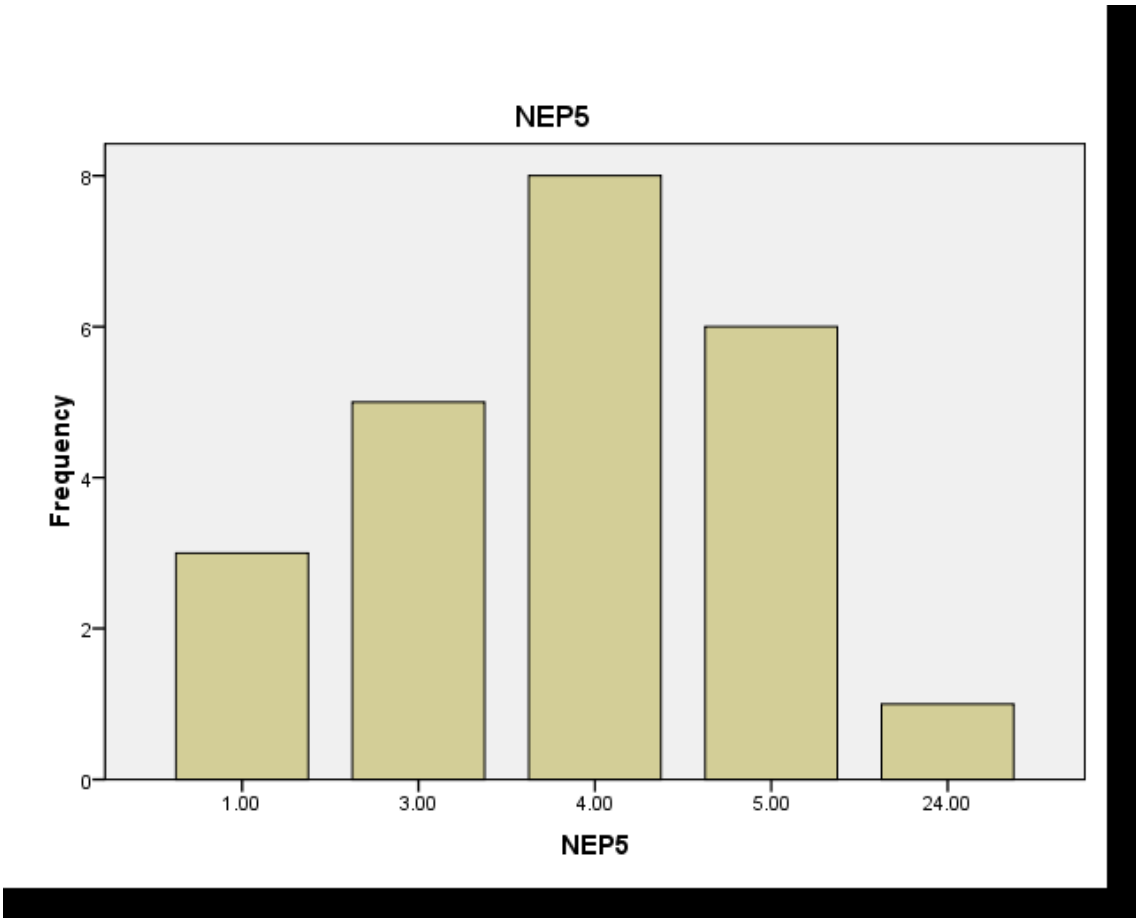

NEP4
Frequency
12
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
11.00
NEP4

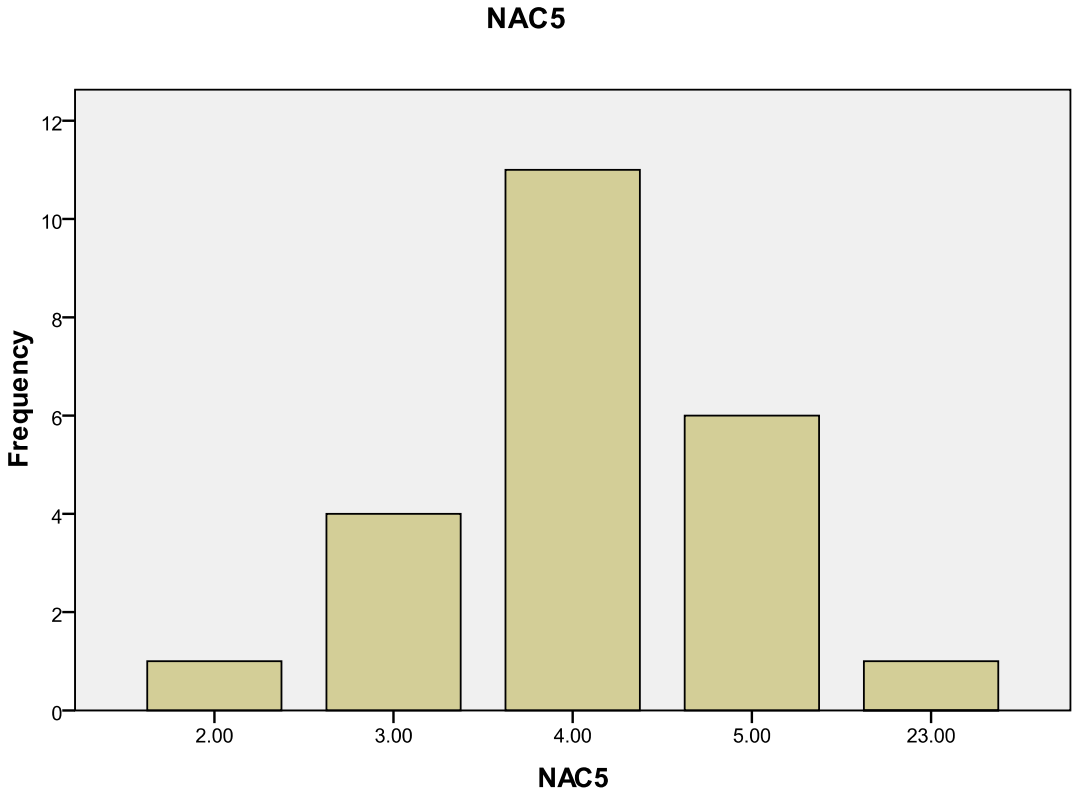

FSM1
Frequency
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
12.00
FSM1

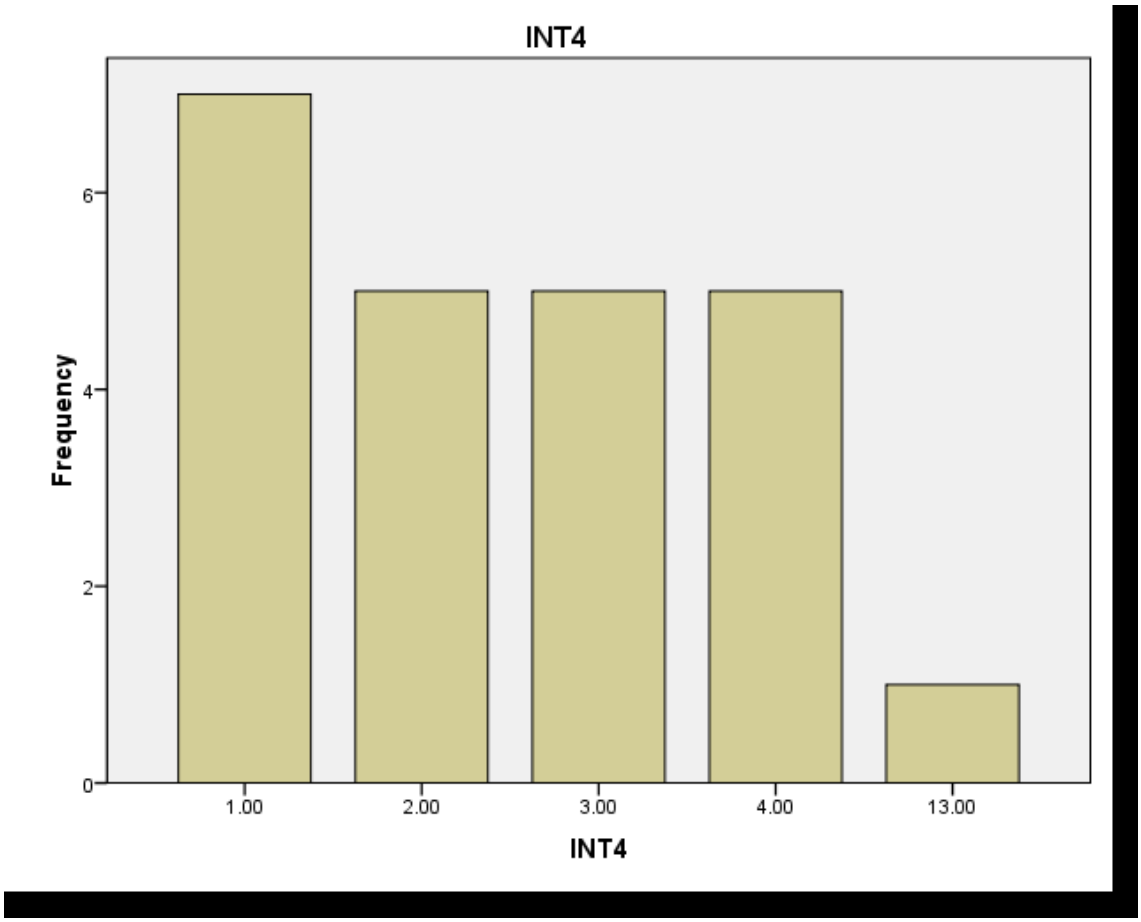

FSM7
Frequency
12
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
20.00
FSM7

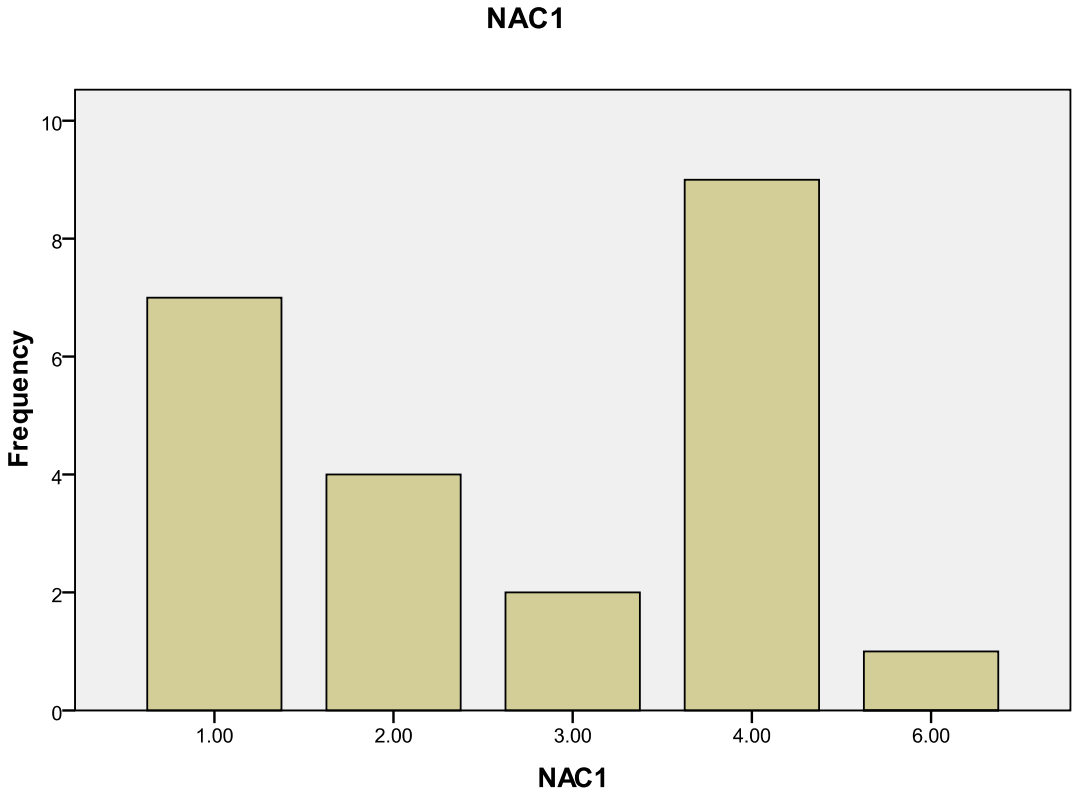

FSM8
Frequency
12
10
8
6
4
2
0
3.00
4.00
5.00
21.00
FSM8


INT6
Frequency
10
8
6
4
2
0
1.00
2.00
3.00
4.00
5.00
22.00
INT6


NEP5
Frequency
8
6
4
2
0
1.00
3.00
4.00
5.00
24.00
NEP5


NAC5
Frequency
12
10
8
6
4
2
0
2.00
3.00
4.00
5.00
23.00
NAC5


INT4
Frequency
6
4
2
0
1.00
2.00
3.00
4.00
13.00
INT4


NAC1
Frequency
10
8
6
4
2
0
1.00
2.00
3.00
4.00
6.00
NAC1