UNIVERSITY OF NEWCASTLE UPON TYNE

Shedding Light on the Predictive Validity of English Proficiency Tests in Predicting Academic Success

Prepared and written by

Laila Whitman Rumsey

Dissertation submitted for the degree of

Doctor of Philosophy - Applied Linguistics

School of Education Communication and Language Sciences
Department of Education

July, 2013

Abstract:

This embedded mixed method case study aims at shedding light on the use of English proficiency exams as placement tests and their viability as predictors of future academic performance. Most predictive validity studies achieve correlation coefficients in the range of 0.20-0.30 (In 1988, Davies suggested that 0.30 was an acceptable correlation for predictive validity studies.) when such exam results are compared with final course marks and/or GPAs, meaning that the results of language proficiency exams tend to have limited usefulness in admissions decisions. The Common Educational Proficiency Assessment (CEPA) is the focus of this research project. The results of the CEPA form a major part of admissions criteria for tertiary institutions in the United Arab Emirates (UAE). The CEPA has been reported to have achieved very high correlations (i.e. 0.699 in 2007) when compared to final first semester marks. This study examined this phenomenon at a large, vocational college. In addition, through an exhaustive exploration of college practices, and the input and opinions of a wide variety of stakeholders, creating a vivid picture of the context within which the CEPA operated, this study identified variables which may contribute to the success of the UAE CEPA as a placement instrument. The efficacy of using regionally-produced exams rather than internationally produced ones to not only gauge linguistic competence, but also to predict future success in an English readiness programme (required before matriculating) is considered. It is hoped that the results of the study may lead to improved predictive validity for regionally or locally produced placement tests at other institutions.

Dedication:

I dedicate this dissertation to my beloved father, Reverend James Rumsey and my cherished grandmother, renowned journalist Ardis Whitman, for what they instilled and inspired in me. I hope that the legacy of who they were lives on in some small way through this accomplishment of mine.

Acknowledgements:

Before the created, I wish to acknowledge my thankfulness to the Creator, Almighty Allah, the Merciful, the Compassionate Who answered my prayers and Whose almost tangible presence in my life has been extremely humbling and motivating at the same time.

There are many people to whom I would like to express my gratitude. First of all, I wish to extend deep feelings of gratitude to my supervisor, Scott Windeatt, for his unfailing and generous support and assistance throughout these often tumultuous seven years. Second, I extend a special debt of gratitude to His Excellency Sheikh Nahayan al-Mubarak and to TIRF for the support of His Excellency and that foundation in the completion of this study. I also thank the students and my former colleagues at HCT, without whose cooperation this research study would not have been possible. Last, but definitely not least, I wish to express my heart-felt thankfulness to my husband, Abdus Sabur, for his encouragement, support and patience – a wonderful partner upon whom I could always depend to raise my spirits and renew in me that sense of purpose. May God bless all of you.

TABLE OF CONTENTS

LIST OF TERMS AND ABBREVIATIONS USED IN THIS STUDY vii

CHAPTER 1: INTRODUCTION

1.1 – PRESENT SITUATION 1

1.2 – ARGUMENTS FOR AND AGAINST THE USE OF SUCH INSTRUMENTS 1

1.3 – THE CURRENT STUDY 2

1.3.1 What is the CEPA?

1.3.2 Historical Overview

1.3.3 The Aims of the Study

1.4 – STUDY OVERVIEW 5

1.5 – SUMMARY 6

CHAPTER 2: LITERATURE REVIEW

2.1 – INTRODUCTION 8

2.2 – CLARIFICATION OF TWO TERMS AND CONSTRUCTS 8

2.3 – VALIDITY 12

2.3.1 – What was being unified, and why?

Criterion-Oriented Validation

Content Validation

Construct Validation

2.3.2 – Validity Unified

2.4 – ISSUES OF RELIABILITY.. 18

2.5 – Validity - The Discussion Continued 20

Kane: Argument-Based Validation

Lissitz & Samuelson: Changing Term. & Validity Emphasis

2.6 – Pesky Predictive Validity 25

2.7 – REFERENCING 28

2.8 – MCQS.. 30

2.9 – SCALED SCORES 32

2.10 – INTRA-RATER, INTER-RATER RELIABILITY AND BAND SCALES 34

2.11 – ISSUES OF PRACTICALITY.. 38

2.12 – TESTING ETHICS: THE CONCEPT OF 'FAIRNESS'.. 40

2.13 – THE AFFECTIVE DOMAIN OF TEST TAKERS 45

2.14 – COMPARABLE STUDIES 46

2.14.1 Recent Predictive Validity Studies

2.14.2 Local/Regional Design & Development of High-Stakes Exams

2.15 – SUMMARY 58

CHAPTER 3: METHODOLOGY

3.1 – INTRODUCTION 60

3.2 – DETERMINING THE FOCUS 61

3.2.1 The Paradigmatic

3.2.2 The Methodological

3.2.3 The Research Method: Case Study

3.3 – DETERMINING WHERE & FROM WHOM THE DATA WILL BE COLLECTED 65

3.3.1 The Institutions Involved

3.3.2 The Research Participants

3.3.3 Research Ethics

3.4 – DETERMINING THE PHASES OF THE RESEARCH 67

3.4.1 Phase One: The Conceptualisation of the Main Study

3.4.2 Phase Two: Qualitative Data Collection

3.4.3 Phase Three: Quantitative Data Collection

3.4.4 Phase Four: Merging the Results

3.4.5 Phase Five: Interpretation

3.5 – DECIDING INSTRUMENTATION 81

3.5.1 Qualitative Data Collected

Questionnaires

Semi-Structured Interviews

Group Interviews

3.5.2 Quantitative Data Collection

3.6 – DATA ANALYSIS PROCEDURE 86

3.7 – PLANNING LOGISTICS 87

3.8 – PLANNING TECHNIQUES TO DEMONSTRATE THE 'TRUSTWORTHINESS' OF THE RESULTS 87

3.9 – SUMMARY 88

CHAPTER 4: RESULTS OF THE RESEARCH

4.1: INTRODUCTION 90

4.2: STUDENT QUESTIONNAIRE RESULTS 90

4.3: SEMI-STRUCTURED INTERVIEW RESULTS 101

4.3.1 Interviews with College Personnel

4.3.2 Interviews with College-Affiliated Personnel

4.4: GROUP INTERVIEWS 132

4.4.1 The First Discussion, December 2007

4.4.2 The Second Discussion, May 2008

4.5: THE ANALYSIS OF THE DOCUMENTATION 140

4.5.1 Pearson's Product Correlation

4.5.2 Multiple Linear Regression

4.6: SUMMARY 150

CHAPTER 5: DISCUSSION AND ANALYSIS

5.1: INTRODUCTION 151

5.2: OVERVIEW OF THE STUDY: THE LITERATURE 151

5.3: OVERVIEW OF THE STUDY: THE RESEARCH DESIGN 154

5.4: RESEARCH QUESTIONS AND FINDINGS 155

5.4.1 Research Question (A) and sub-questions

5.4.2 Research Question (B) and sub-questions

5.5: IMPLICATIONS FOR THE DESIGN AND DEVELOPMENT OF REGIONALLY OR LOCALLY PRODUCED PROFICIENCY EXAMS 172

5.6: THE THESIS QUESTION 173

5.7: SUMMARY: THE SIGNIFICANCE OF THE FINDINGS 173

CHAPTER 6: CONCLUSION

6.1: INTRODUCTION 176

6.2: IMPLICATIONS FOR THEORY 176

6.3: IMPLICATIONS FOR PRACTISE 177

6.4: LIMITATIONS 180

6.5: IMPLICATIONS FOR FUTURE RESEARCH 183

6.6: SUMMARY 185

BIBLIOGRAPHY 186-192

LIST OF GRAPHICS (TABLES, FIGURES, DIAGRAMS, INSERTS, CHART)
TABLES

TABLE 2.1: TAXONOMY OF TEST EVOLUTION PROCEDURES 22

TABLE 2.2: ADVANTAGES & DISADVANTAGES OF NORM & CRITERION REFERENCED EXAMS 29

TABLE 2.3: SOME PREDICTIVE VALIDITY STUDIES 50

TABLE 3.1: PHILOSOPHICAL DIMENSIONS OF PRAGMATISM 63

TABLE 3.2: RESEARCH METHODS GRID 68

TABLE 4.1: STUDENTS' PERSONAL DETAILS FROM THEIR QUESTIONNAIRE RESPONSES 92

TABLE 4.2: SELECTED STUDENT COMMENTS FROM OPEN RESPONSE PART OF ITEM #21 96

TABLE 4.3: SELECTED STUDENT COMMENTS FROM OPEN RESPONSE PART OF ITEM #30 99

TABLE 4.4: CEPA ENGLISH SCORES VS. IELTS AND TOEFL SCORES 130

TABLE 4.5: PEARSON'S CORRELATION FOR DIPLOMA FOUNDATIONS (DF) 143

Table 4.6: Pearson's Correlation for Higher Diploma Foundations (HD) 144

Table 4.7: DF Results of the MLR Analysis of Variables with 10% of Sample 149

Table 4.8: HD Results of the MLR Analysis of Variables with 10% of Sample 150

Table 5.1: Facts from Studies of Regionally-Developed English Proficiency Exams 171

Figures and Chart

Figure 2.1: Lumley's Model of the Rating Process 38

Figure 3.1: Diagram of the Research Study's Convergent Sequential Design 80

Chart 4.1: Responses to Questions 18-21, DF and HD 94

Diagrams and Inserts

Diagram 4.1: Pearsons Correlation Scatterplot Diagram of DF Analysis 143

Diagram 4.2: Pearsons Correlation Scatterplot Diagram of HD Analysis...................... 144

Diagram 4.3: Excel Representation of the Scanned Questionnaire Responses 146

Insert 3.1: Student Questionnaire 69-76

Insert 3.2: Semi-Structured Interview Protocol..77-78

Insert 4.1: 0155 (DF): 90% MLR Stepwise Results 148-149

APPENDICES (on separate disk)

Appendix A: Information on Quoted Predictive Validity Studies

Appendix B: Student Questionnaire

Appendix C: Questionnaire Pilot Results

Appendix D: Tabulated Student Questionnaire Results

Appendix E: Interview Protocol

Appendix F: Transcribed Interviews

Appendix G: Transcribed Group Discussions

Appendix H: Multiple Linear Regression Output

Appendix I: Permission Forms for Conducting Research

Appendix K: Official Translation of Questionnaire

Appendix L: College KCA Reports (w/Correlations) for AY 2005/6, HD and DF

Appendix M: CEPA Sample Exams

Appendix N: CEPA Training Workshop Document

List of Terms and Abbreviations used in this Study

assessment: primarily refers to formal, summative test or exam in this study

CEPA: *Common Educational Proficiency Assessment.* The CEPA office runs the CEPA English and CEPA Maths exams, and a prof'l development program for Ministry of Education English teachers. CEPA English is used to determine eligibility for selection into degree and higher diploma courses at UAE University, Zayed University and the Higher Colleges of Technology. From 2007 was administered to all Grade 12 students completing the Ministry of Education's English curriculum. CEPA is also used to determine placement into preparatory programs at the three institutions (CEPA website, 2012).

DF: *Diploma Foundations.* College preparatory track for students admitted to college with fairly poor English ability

embedded case study methodology: a way of integrating quantitative and qualitative methods into a single research study in case studies containing more than one sub-unit of analysis (Scholz & Tietje, 2002; Yin 2003)

exam/ test: used synonymously

GSC: *General Secondary Certificate.* Equivalent to high school diploma, it is a total score of all the final marks of the 12th grade subjects

HCT: *Higher Colleges of Technology.* A UAE government subsidized and supervised vocational college. Admission is restricted to UAE nationals & nationals of GCC countries

HD: *Higher Diploma Foundations.* College preparatory track for students admitted to college with fair English ability

high school/ secondary: used synonymously

KCA: *Key Common Assessments.* A euphemism for the final exams of the foundations courses at HCT

MCQs: multiple choice questions

MM: mixed method research (i.e. quantitative and qualitative)

NAPO: The *National Admissions and Placement Office* of the UAE. It "was established in 1996 to coordinate all applications and admission of UAE nationals to the United Arab Emirates University (UAEU), the Higher Colleges of Technology (HCT) and Zayed University (ZU). In addition, NAPO is responsible for developing and implementing bridge programs between secondary and post-secondary education. NAPO provides advice to higher education institutions about application and enrollment trends, and oversees the Common Educational Proficiency Assessment (CEPA)" (from NAPO website, 2012).

NNS: non-native speaker (of a language)

PI: *Placement Index.* A formula used by the Registrar at HCT for admissions decisions

1: INTRODUCTION

1.1 Present Situation

When one reviews the statements made by the major players in the market of international, standardized, criterion-referenced examinations of proficiency in English, such as TOEFL, IELTS and even GMAT, one is immediately struck with the observation of an apparent contradiction between what these organisations state should be the intended use of the results of these examinations, and what, in reality, actually happens. In fact there is what seems to be some incongruence in their own explanations of purpose. For example, the Educational Testing Service (ETS), the organisation which produces TOEFL, clearly states that "the TOEFL test is a measure of general English proficiency" (2001:18) IELTS also clearly states that theirs "is a highly dependable, practical and valid English language assessment" (2005). However, both organisations acknowledge the use of their exam results as an admissions requisite: "TOEFL test scores help determine whether an applicant has attained sufficient [English language] proficiency … to study at college or university" (2001:18), and IELTS states that their exam is "primarily used by those seeking an international education…" (2005). GMAT states that its "scores cannot be used to estimate potential for success in a career [nor can it] pinpoint achievement in specific subjects…", but in another section of its website states, "In repeated research studies, GMAT scores have been found to be a good predictor of academic success in the first year of an MBA…" (2005) Essentially these organisations are saying on the one hand that their exams assess English (or business) competence only and cannot or must not be used to predict academic performance, but on the other hand, imply the opposite when they say that their results are, or can be used as, admission criteria. The research to support this sort of usage is inconclusive. What specific research basis they have for taking either position is not mentioned in the readily-accessible literature that these international standardized examination-producing organisations provide. In general, what has been written on this topic would seem to suggest that this is not a valid use.

1.2 Arguments for and against the Use of such Instruments

Well-known English proficiency exams, most notably the TOEFL (produced by Educational Testing Services- ETS) and the IELTS (produced by Cambridge), have established their credibility and reliability internationally through rigourous research, development and periodic re-evaluation. This has led to a situation where the authorities who make decisions about which measure of English competence they wish to adopt rely on these measures without seriously questioning the relevance of all the uses to which they are put. Based on their research of an English readiness programme at a university in the Persian Gulf region, Davidson and Dalton stated that "achievement of a certain degree of language proficiency, as determined by an international benchmark such as TOEFL may be a necessary, but it is not a sufficient condition on which to base decisions regarding students' readiness to enter into baccalaureate study" (2003:45). Speaking on the use of IELTS as a measure of competence in English in Germany, Patrick Griffin asserted that "reliance on a single cut-point and a dichotomy of competency is unworkable and ignores errors of measurement and variability" (2001:98). IELTS itself has, since 1995, been involved in an extensive revision project, which has produced some valuable research insights into its predictive validity. In their paper (Paper 4), Fiona Cotton and Frank Conrow discovered that there were "no positive correlations found between IELTS scores and [English] language difficulties students reported with aspects of their coursework. Qualitative data indicate that language difficulties are one of many variables affecting academic achievement."(2004) Even so, in Paper 5, written by Clare McDowell and Brent Merrylees, they say "Many universities run their own English test for entry into programs possibly as an incentive to lure students to their university. Other institutions may use IELTS but with little understanding of what an IELTS score actually signifies and what level of predictive validity it offers." (2004) Noble, Schiel and Sawyer have said that this kind of use can be valid, provided that the course placement system is systematically evaluated, but even they acknowledge the inherent difficulty of allocating resources for this purpose: "the resulting decisions [about the evaluation of the placement system] are often difficult because the required resources may be substantial and could be allocated to other worthy programs or projects" (2003:12).

1.3 The Current Study

1.3.1 What is the CEPA?

This is the acronym for an English proficiency test named the Common Educational Proficiency Assessment or "CEPA". It was developed in the United Arab Emirates (UAE). It is one of the requirements for admission to tertiary education in the UAE at any one of the three government subsidized universities. Originally a low-stakes assessment, CEPA is now a high-stakes exam for which a certain level must be reached in order not only to gain admission to tertiary education, but also to exit secondary education (high school). Its results are primarily used for placing students in the most appropriate pre-sessional foundation level at government tertiary institutions in the UAE. (The purpose of these "foundation" courses is to ensure that students entering tertiary institutions with limited English, Maths and computing skills are brought up to a level where they can cope more easily with the demands of degree or diploma programmes at these institutions.) At the Higher Colleges of Technology, these foundation courses last a year. The CEPA consists of three parts; Grammar, Reading and Writing. The test duration is 2 hours and 15 minutes. There is no break time between the three parts of the exam.

1.3.2 Historical Overview

CEPA was developed originally to facilitate the placement of students for English language study purposes for the tertiary educational institutions of the UAE at the beginning of this century. From 2006, CEPA-English has been used to determine eligibility for admission and placement into pre-Bachelors, academic readiness courses, named "Foundations" and Bachelors degree courses, and thus became a so-called "high stakes" examination. This study focused on a branch of one of these three institutions: the Higher Colleges of Technology. There are two pre-Bachelors foundations programmes: Higher Diploma (HD) and Diploma Foundations (DF). In 2007, applicants had to achieve a minimum score of 170 on CEPA English, in addition to a minimum average of 70% on the General Secondary Certificate (GSC) to be eligible for placement into the Higher Diploma programme at the HCT. (The GSC is a composite score of all the final marks of students' high school courses.) Those scoring less than 160 on the CEPA-English were automatically placed in HCT's Diploma Foundations programme.

1.3.3 The Aims of the Study

The research sought to answer several basic questions about predictive validity, including the question of whether high predictive validity for first-year academic performance could be determined for this English proficiency exam (the CEPA) which is used as a selection tool amongst other uses, and whether it is possible to identify other variables that may enhance or ensure academic success in college-level courses. The CEPA has been reported to have achieved fairly high correlations with students' final English marks (i.e. r=.699 in 2007) at the end of a one-year academic foundations course of study, which precedes formal admission to the college. (Correlations with other subjects were either not investigated by the college or not reported.) It was the aim of the research to closely examine and explore this phenomenon, and to identify as many other variables as possible which might contribute to the aforementioned relatively high predictive validity of the UAE CEPA as a selection and placement instrument in a large, vocational college. This aim was critical in the establishment and demonstration of the importance of a thorough understanding of the context in which an assessment is being used. The application of embedded case study methodology was very useful in enabling the complete investigation of the situational context of the research study.

Predictive validity studies have come under a great deal of criticism for several reasons, but two points in particular have raised the most serious concerns. First, there is the issue of the 'truncated sample'. When the researcher sets out to compare entrance scores and scores at the end of a course, he or she has only the cohort of students who were admitted and would obviously not be able to include those who were not admitted, rendering a true estimate of the predictive validity of the selection tool questionable. However, in 2007 an unusual situation occurred at a particular branch of this college, which enabled the college to admit virtually every student who applied, what Prof. Charles Alderson whimsically referred to as a "virgin cohort" (personal discussion, August, 2005).

Second, the student population of most predictive studies has typically been one of great variance in educational experience and personal background. This makes determining predisposing variables of academic success extremely difficult. The students in this cohort are a fairly homogenous group (i.e. all male Emirati citizens between the ages of 17 and 21). The uniformity of their educational experience and backgrounds facilitated the identification of several variables which appeared to positively correlate with academic success.

The purpose of initiating this research study was essentially to investigate three premises:

1. High, credible predictive validity is possible and is facilitated by a thorough investigation, and subsequent detailed description, of the context in which an assessment (or assessment scheme) is being administered.
2. Within a homogenous cohort of students that was offered almost unfettered admission to a regional college, it is possible to identify some variables that positively correlate with academic success.

3. Regionally-developed English proficiency examinations can be a dependable, reliable, and perhaps even a preferable alternative to international examinations when used for regional institutions.

1.4 Study Overview

The thesis question which guided the research was: "Is strong predictive validity possible for regionally-developed English proficiency examinations used for admissions and/or placement decisions?" The main focus of study in this research project was the CEPA exam. The research was guided by the premises under investigation through the following questions:
A) Is this locally-produced English proficiency examination a "good" predictor of English competence?

Two of the sub-questions cover clarifications about the validity and reliability of the CEPA itself, and its predictive validity in gauging the ability of students to progress successfully in English courses. How CEPA scores figure into the admissions process will also be considered.

Sub-Questions:

1) Is the CEPA study a valid and reliable estimate of students' English ability before admission?
2) How are the results of the CEPA exam used in the college's admissions and placement decisions?
3) What is the predictive validity of this English proficiency exam in gauging the ability of students to progress successfully in English courses and Maths courses in which the medium of instruction is English?

B) What variables appear to positively correlate with students' ability to succeed academically? (The word "factors" is interchangeable with "independent variables", and will be used as such in this research. See Field, 2005: 731.)

The three sub-questions to this question explore issues concerning personal, motivational and attitudinal information about the student cohort.

Sub-Questions:

1) Are the ages and academic levels of the research participants homogenous?
2) Is the socio-economic background of the participants homogenous?

3) What are their motivations for, and attitudes towards, learning English of the study's participant students?

The subjects of the investigation included the entire cohort of first year "Foundations" students at a campus of the Higher Colleges of Technology, a vocational college in the United Arab Emirates. They numbered approximately 350 students. The cohort consisted of all male, Arabic-speaking Emirati nationals, primarily between the ages of 17 and 25. The researcher was also granted access to the admissions files of the first-year, foundations student cohort, as well as to their academic records at the college for the first year. This obviously facilitated the design and implementation of this mixed method (MM) research project.

This project may also be described as a convergent mixed method study since data obtained from both qualitative and quantitative sources were first examined separately, and then merged finally in analysis. In order to achieve a detailed and thorough accounting of the context of the study, an embedded case study design was employed. As pointed out by Yin (2003: 8), the embedded case study is particularly relevant to the examination of a situation where the relationship between the phenomenon of interest and context are not clearly evident. Data gathered included students' responses to a questionnaire. The questionnaire itself went through several stages of revision, not the least of which was the pilot of the instrument with off-sequence students of the previous cohort. There were also two follow-up group discussions with students of this cohort at mid-year and at the end of the academic year to discuss any arising matters of concern with the students and to clarify issues raised in the free-response section of the questionnaire itself. College administrators and teachers were also interviewed in order to have as many perspectives as possible within the college as to the factors which may enhance students' chances for academic success, as well as whatever extrinsic and intrinsic motivations students may have for continuing their education, and whatever hindrances they may also face. This of course, is in addition to the data about the students which was supplied by the college.

1.5 Summary

In Chapter 2, the literature relevant to the study is examined. The chapter has been broadly organised into four major topic areas: 1) an overview of the ongoing discussion of the terms validity, reliability and practicality within the field of applied linguistics and testing in general; 2) an explanation of test item types and scoring systems that are a part of CEPA; 3) an overview of test design issues, including testing ethics and the affective domain of test-takers; and finally, 3) an overview of other predictive validity studies and articles about locally- or regionally-developed English proficiency exams.

In Chapter 3, the research design and methodology is presented. The chapter begins with a consideration of where the research design stands with regards to the more overarching focuses of paradigm and methodological decisions. Then it proceeds to detail the phases of the research, including the data collected, site selection and access negotiation, and what procedures were followed to assure, as much as possible, the trustworthiness of the study's results.

In Chapter 4, the results of the research are detailed, starting with the qualitative instruments, and how the data were collated and organised. Then, an explanation of how the quantitative data was analysed and the rationale for the analyses follows, concluding with the results of the Multiple Linear Regression, which combined data from qualitative and quantitative sources. Chapter 5 discusses and analyses the results, focusing on the original research questions. It includes the study's major findings, as well as reflections on their significance. Implications for theory (in terms of practical application), practice and future research are covered in Chapter 6.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Within the field of education, most experts concur about the necessity of some sort of testing, upon the condition that it should not be the only method for assessing a student's language ability (Stobart, 2008; Black and William, 1998; Sadler, 1989; Wiggins, 1997; Biggs, 1998; Klenowski, 1996; Broadfoot in Nuttall, ed., 1986; etc.). These experts have pointed to the need for a 'more balanced' approach to the way we assess our students' language abilities in that many factors need to be and should be considered, rather than an over-reliance on one assessment instrument. Balancing an assessment scheme should produce a multi-dimensional view of a student's capabilities in much the same way triangulation in research aims to be comprehensive by combining operational techniques. But this avenue is fraught with pitfalls as well, not the least of which is the terminology itself. If we seek to make research in Applied Linguistics more scientific, and therefore more accessible and comparable to other research studies, then it is absolutely necessary to establish a uniformity of terminology, at least as concerns the major concepts of the field such as 'assessment' and 'testing', and 'competence' and 'proficiency'.

2.2 Clarification of Two Terms and Constructs

There is much variance in the most fundamental of terms used by assessment experts and writers, and by others centrally involved in assessment processes. For example, Lyman's book, *Test Scores and What They Mean* (1991), offers detailed and exhaustive explanations of, and intended uses for, the different kinds of standardized, norm-referenced tests and their components. Still, he never defines the actual word *test*. A review of different dictionaries of educational terms reveals quite a range of different, and sometimes confusing, explanations of the word 'test' (*The International Dictionary of Education* (1977), *A Critical Dictionary of Educational Concepts* (1986), and *A Dictionary of Education* (1959) which sub-lists no less than 340 different uses for the word 'test'!).

Ur defined the word 'test' "as an activity whose main purpose is to convey (usually to the tester) how well the [test taker] knows or can do something. The test gives a score which is assumed to define the level of knowledge of the test taker," (1996: 33) but she did not clarify how this might differ from 'examinations'. There is no definition for the word 'test' or 'examination' offered in Bachman and Palmer's *Language Testing in Practice* (1996/2009), yet both words are included. Perhaps this is because Bachman devoted space in his earlier publication, *Fundamental Considerations in Language Testing* (1990), to an explanation of the difference. In clarifying the difference between the terms 'evaluation', 'test' and 'measure', he reached back to quote Carroll's (1968) definition of a test as a "… procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual." (Carroll, 1968: 46, quoted in Bachman, 1990: 20) Later, in the 'Notes' section of the same chapter, Bachman (1990: 50) observed that the "distinction is sometimes made between 'examinations' and 'tests'". And he went on to say, "But as Pilliner (1968) pointed out, there is no consensus on what the distinction is." Because of this, Bachman stated that he considers the terms 'examination' and 'test' "stylistic variants of 'evaluation' and 'test'" and then did not consider the matter further. McNamara echoed the seeming inter-changeability of these two terms, 'examination' and 'test' when he said, "Paper-and-pencil tests take the form of the familiar examination question paper" (2000: 5). This sentiment is also supported in the *Longman Dictionary of Language Teaching and Applied Linguistics* (Richards and Schmidt, 2002: 189), which states: "The terms 'examination' and 'test' can be used interchangeably as there seems to be no generally agreed-upon agreement regarding the distinction between the two."

In a guide written for laypeople, Moon (1991: 25) spells out what he sees as the distinction between these terms: "Tests now tend to be seen as limited activities, perhaps involving spelling or mental arithmetic that contribute to the ongoing process of assessing. Examinations are usually seen as final, marking the end of the year or the end of the school process." And this seems to be the general consensus amongst professionals involved in educational assessment.

For the purpose of this study, I will rely on McNamara's (2000: 11) definition of *test*: "A language test is a procedure for gathering evidence of general or specific language abilities from performance on tasks designed to provide a basis for predictions about an individual's use of those abilities in real world contexts." This definition is most appropriate here because the focus of this study is upon the predictive usefulness of the inferences that are made about the results of the CEPA examinations in the U.A.E. (See page 4 for a list of terms and abbreviations which are specific and/or relevant to this study.)

Another terminological difficulty which exists within the field of testing is between the constructs of 'competence' and 'proficiency' when describing language ability – and this goes right to a core issue in the field of Applied Linguistics. As stated by Fulcher and Davidson (2007:36), "In language testing and assessment, we have come to rely on models of what is variously called language proficiency, communicative competence or communicative language ability (CLA). The terminology that is used has become confusing". Chomsky, in his pivotal text, *Aspects of the Theory of Syntax* (1965: 3-5), described 'competence' in terms of language competence, meaning the speaker-hearer's knowledge of his language, and language performance, meaning the actual use of language in concrete situations. In a desire to further clarify Chomsky's notion of language competence, Hymes (1966: 115) further defined language competence by coining the term "communicative competence". His ethnographic exploration of communicative competence included "communicative form and function in integral relation to each other" (Leung, 2005: 122). Chomsky's descriptions were echoed by Dale (1983: 64), when he proposed that competence consisted of "linguistic competence" (i.e. knowing the rules of the language), and "linguistic performance" (i.e. putting this knowledge of a language into action). In the same year, Stern (1983: 342-343) put forward two "aspects of competence": the first being that knowledge of the rules governing a language and the ability to apply these rules "without paying attention to them", the second aspect being that "intuitive grasp of the linguistic, cognitive, affective and sociocultural meanings expressed by language forms". In these quoted texts, neither Dale nor Stern spoke of language ability in terms of 'proficiency'. Harmer (1991: 18) also further developed the notion of communicative competence by stating that language speakers have a "subconscious knowledge of the rules [and this] allows them to produce an infinite number of sentences", and he separated the construct of competence into communicative competence and strategic competence. The latter he described as "knowing how to use language rather than just knowing about language" (Harmer, 1991: 19). After explaining the competence/control model proposed by Bialystok and Sherwood-Smith (in which competence is viewed in terms of knowing one's language system and control is viewed in terms of being able to use that language knowledge to communicate effectively), Cook (1996: 164) said that "learning a second language involves two things: the knowledge that makes up competence and the control that is used in producing speech".

These scholars have largely focused on describing language ability in terms of competence, so what about 'proficiency'? In a recent email exchange on the L-TestL forum (June 23, 2013), Professor Tim McNamara said that he thinks that the terms competence and proficiency "have different histories and overlap in meaning, but [that] there are differences". McNamara continued, saying, "Competence is the term introduced by Chomsky and picked up by Hymes, which became one of the bases for communicative language testing (Canale and Swain 1980, Bachman 1990, etc.), so that from the communicative movement onwards it became a widely used alternative to the term proficiency". McNamara (1996: 6-35) devoted a chapter in a book he authored, *Measuring Second Language Performance*, to exploring the contributions of applied linguists to language learning theories and language testing models in the 20th century. In that chapter, 'competence' was invariably a characteristic of language acquisition theory, and 'proficiency' was invariably used with regard to test description. Llurda (2000: 85-95) presented perhaps the most thorough attempt to unpack and explain the difference between 'competence' and 'proficiency'. In his article, Llurda (2000: 85) essentially limits the term competence to how it is understood in Chomskyan terms, "whereas 'communicative language ability' ought to be applied to speakers' ability to use a language, [which should be further] divided into two components, namely *language proficiency* and *communicative proficiency*" (italics in the original). Finally, Professor Hamp-Lyons seems to equate the term 'performance' to 'proficiency when she described her understanding of the essential difference between 'competence' and 'proficiency' in a post to L-TestL, "I was taught that competence - performance were a pair like the Saussurian langue - parole. In a way the langue/parole distinction is easier for language people like us because we can't mix them up with other close cousins. Langue is the potential system that is available in a given society; parole is the actualisation of that system by the individuals of that society. …This is very similar to (or the same as, maybe) competence - performance: Competence is the mental set of linguistic representations that we acquire through immersion in a particular speech community. Performance is the use of that knowledge by individual members of the community. This might explain why we often find persons (including plenty of native speakers) who are quite 'competent' in English within prescribed contexts, but become incompetent in other contexts. Competence is happening in the mind, it's cognitive, while performance is taking place in the social world; it's interactive and interpersonal. Saussure said that language = langue + parole. I guess that means that to be truly "proficient" we need both" (L-TestL post, 24/6/2013).

After reviewing the statements and explanations of several language learning and testing experts and scholars, it appears that 'competence' refers to knowing the rules of a language and how to appropriately use them in communication, while ‘proficiency’ has used to describe a type of language competence, or used as a term that refers to a kind of assessment instrument that aims to measure examinees’ degree of language ability. As pointed out by Jones (L-TestL post, 25/6/2013), proficiency “does not exist until someone measures it. This distinguishes it from other terms such as "ability" or "competence" which refer to properties of learners which exist independently of whether they are measured or not”. It is this understanding of proficiency that will be applied in this study.

2.3 Validity

Validity has been one of the most intensely discussed topics amongst testers and applied linguists. The validity of an assessment instrument is a critical quality of a test, because it is a gauge of its ‘test-worthiness’; of its appropriateness for the use/s to which it is being put. One gets a sense of the how intense the debate about this amongst applied linguists became in the 1980s in this quote from Cronbach (1988:3): “Validation was once a priestly mystery, a ritual performed behind the scenes, with professional elite as witness and judge. Today, it is a public spectacle combining the attractions of chess and mud wrestling. Disputes about the appropriateness of tests impose a large responsibility on validators”. This statement reflects the tremendous transformation of test validation from one in which ‘experts’’ opinions were unquestioningly accepted at face value, to the situation today where validation is scrutinized by a large body of test stakeholders, and about which the test validators are being held accountable.

Validity as a concept has not changed much since Garrett (1937) offered this definition: “The validity of a test is the fidelity with which it measures what it purports to measure” (p.324, quoted in Angoff, 1988: 27). About validity, 63 years later Bachman (2000: 23) said that “it is the fundamental basis upon which considerations of values, uses and consequences are based”. Establishing the validity of an assessment has always been a primary concern, but the way in which we do this has undergone several evolutions since Garrett’s time.

Messick, in his seminal article on validity (1989: 13), declared that validity is fundamentally a single umbrella issue involving and including other sub-issues. In this article, he put forward his idea of a unified theory of validity. Moss (2007: 470) described the concept of unified validity as a reflection of "the generative vision of scientific inquiry". And about Messick's theory, Keith Markus (1998:7) said, "[It] is profoundly influential in part because it brings together disparate contributions into a unified framework for building validity arguments". What have been these 'disparate contributions'? "An understanding of the unitary concept of validity requires an understanding of what, in fact, was being unified and why" (Moss, 2007: 471).

2.3.1 What was being unified and why?

Before the 1950s, test-criterion correlations were the standard for judging test accuracy. "In the 1940s, this identity was so strong that the term validity came to be used synonymously to mean a predictive correlation coefficient." (Shepard, 1993: 410). In what was perhaps the first major impetus to credit validity with much greater importance than it had previously been given, Cronbach and Meehl (1955: 281) classified validity into distinct categories in a paper emanating from their work on the first codification of validity standards for the American Psychological Association (APA). The named categories were (1) concurrent, (2) predictive, (3) content and (4) construct, but even within this same paper, the authors pointed out that "the first two of these may be considered together as *criterion-oriented* validation procedures". Later on, Guion (1980: 384) labeled this "the trinitarian view" in what Wood (1993: 146) described as the "churchy language of the day". These categories have withstood the test of time and remain a focus of discussion. What do they include?

Criterion-Oriented Validation:

Cronbach and Meehl (1955: 282) were very clear about what they believed should be understood by this type of validation: "The investigator is primarily interested in some criterion which he wishes to predict. He administers the test, obtains an independent criterion measure on the same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, he is studying *predictive validity*. If the test score and criterion score are determined at essentially the same time, he is studying *concurrent validity*" [italicised words in the original]. Predictive validity will be discussed more thoroughly later (See pages 34-37.). Both predictive and concurrent validities rely upon some sort of exterior criterion with which to compare the results of an assessment. If there is a good correlation between the results and an established criterion for that assessment, it is said to have good concurrent validity. In this respect, concurrent validity strays close to the realm of reliability, since it is not concerned with the assessment itself, but of the degree of correlation between it and an established criterion.

Bachman (1990: 248) declared that “concurrent criterion relatedness is undoubtedly the most commonly used in language testing”. Major limitations of this kind of validity include the assumption that the criterion itself is valid, and the need to establish that the results of the assessment are not related to measurements of other abilities.

Content Validation:

Starting again with Cronbach and Meehl’s (1955: 282) explanation of validation categories, they stated that content validation “is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test”. Content validity concerns how relevant the assessment is vis á vis the syllabus for the subject being tested. Cronbach (1971: 444) described the evaluation of content validity as “showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn”. Hughes (1989: 22) noted that the greater a test’s content validity, the more likely it is to accurately measure what it is supposed to measure. However, content validation has been criticised because variables such as students’ reading ability can confound efforts to properly define an assessment’s content validity, especially if a limited knowledge of the language is interfering with the student’s ability to demonstrate his or her understanding of the syllabus’ content. Therefore, whether the subject’s content or the student’s ability (innate or learned) is actually being assessed is often difficult to ascertain. High content validity means that the items are a reasonable, relevant and representative example of the material being tested.

Construct Validation:

Cronbach and Meehl (1955: 290) define construct validity in the following manner, "Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim". In this definition, the focus is on theory building or theory testing. Bachman quotes Messick (1975: 957, in Bachman 1990: 255) as saying: "A measure estimates how much of something an individual displays or possesses. The basic question of construct validation is: What is the nature of that something?" Palmer and Groot (1981: 4) view construct validation as a theory testing procedure and distinguish it from all types of validity in which reference to a criterion is important. In their definition, the importance of exploratory factor analysis and confirmatory factor analysis is emphasized. They maintain that in construct validation, one validates a test not against a criterion or another test, but against a theory. To investigate construct validity, one develops or adopts a theory which one uses as a provisional explanation of test scores until, during the procedure, the theory is either supported or falsified by the results of testing the hypotheses derived from it. Even though this points to a process, they may be advocating an over-simplified process. As Stevenson (1981: 50) has said: "Because constructs are being dealt with, there is no clear-cut procedure which would yield a yes-it-does or no-it-doesn't answer. Rather a logical orientation is involved, and it is basic to the concept that it *not* be identified with any [single] investigative procedure". Angoff (1988: 26) defines construct validity "as a process, not a procedure, [requiring] many lines of evidence, not all of them quantitative" and he adds that in order to establish construct validity, it's not enough to merely rationalise or theorise about "the nature of a construct". These are two interesting points. One is that validation is a matter of collecting evidence from many lines of inquiry. The other point is that validation is a process that should take into account practical issues. This sentiment is also echoed by Shepard (1993: 411) who said, "Empirical relations are necessary but not sufficient to establish the validity of test use". This is a logical progression from what Cronbach and Meehl originally stated in 1955:291 as necessary in order to establish a test as a measure of a construct: "A rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct. To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist". Fulcher and Davidson (2007: 8) simplify what is meant by this net saying that it basically contains two simple things: "Firstly, it contains a number of constructs, and their names are abstract, like … 'fluency' and 'accuracy'… Secondly, [it] contains the observable variables – those things that we can see and measure directly". They go on to say that the nomological "network is created by asking what we expect the relationship between 'fluency' and 'accuracy' [for example] to be… Stating this kind of relationship between constructs therefore constitutes a theory and theory is very powerful" (ibid.). Because of the comprehensive nature of the *process* involved, construct validity has the potential to be the strongest and most thorough type of test validity. This is most clearly represented in Messick's 'unified theory of validity'.

2.3.2 Validity Unified

In 1989, about validity Messick (1989: 20) wrote, "Traditional ways of cutting and combining evidence of validity, as we have seen, have led to three major categories of evidence: content-related, criterion-related, and construct-related. However, because content- and criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence". Four years later, Shepard (1993: 423) spoke about Messick's concept as "cement[ing] the concept" that construct validity should be viewed as an umbrella term unifying all the types of validity heretofore regarded as separate entities, "extend[ing] the boundaries of validity beyond test score meaning to include relevance and utility, value implications and social consequences". Again, in the consideration of different types of evidence, Fulcher and Davidson (2007: 12) noted that Messick's unified validity framework facilitates how these types of evidence "contribute in their own way to our understanding of construct validity."

Messick (1998: 41) defined construct validity in the following way: "Construct validation embraces all of the statistical, experimental, rational, and rhetorical methods of marshalling evidence to support the inference that observed consistency in test performance has circumscribed meaning." For example, to understand whether a piece of research has construct validity, three steps should be followed. First, the theoretical relationships must be specified. Second, the empirical relationships between the measures of the concepts must be examined. Third, the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure being tested (Carmines & Zeller, 1991: 23). In this way, as we examine the construct to be measured by the test, "we cause a continuing research interplay to take place between the scores earned on the test and the theory underlying the construct. Viewed this way, we see that all data that flow from the theory, including concurrent and predictive data – but not exclusively so – are useful for [establishing] construct validity" (Wainer, 1988: 26). So, we can say that "'validity' is not a property of a test or assessment, but the degree to which we are justified in making an inference to a construct from a test score… and whether any decisions we might make on the basis of the score are justifiable." (Fulcher and Davidson, 2007: 12) This is why Messick (1975; quoted in Wainer, 1988: 27) declared that *content validity* is in fact, not a kind of validity at all in the sense shared by the other types, or aspects, of construct validity, since content validity is determined by a procedure directly related to the test or assessment itself and not extrapolated from the results. In contrast, the determination of construct validity is a *process* which requires many different sources of evidence, quantitative as well as qualitative. Therefore, construct validity is not simply a property of a measure but is a reflection of and resides in the conditions of its use.

As stated by Bachman (2000: 21), Messick's unitary view of validity has become a valuable basis upon which language testing researchers may consider issues of test result interpretations, as well as the consequences of test use. Years later, this view seems to have gained even greater acceptance, as reflected in Fulcher and Davidson (2007: 14) when they declared that "Messick's way of looking at validity has become the accepted paradigm in psychological, educational and language testing". The APA guidelines for educational and psychological testing for 1985, 1999, and 2006 all state that validity is a unitary concept. In 2006, the APA (2006: 9) stated that "although, evidence may be accumulated in many ways, validity refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself". In this definition the focus is on test scores, not the test itself, as the object of the validation process and enquiry. Referring back to the discussion of content validity, it would seem that unitary validity and the APA, as made clear in their definitions, have rendered the consideration of content validity as almost irrelevant. Additionally, is it the case that Messick's unitary view of validity did not include the test itself? Kane (2007: 77) stated that "[t]he conception of validity as stated by Messick (1989) pulled together all of the strands traditionally included under the heading of validity and added (or at least emphasized) some new strands. Messick organised this large body of material within a very broad conceptual framework involving two interconnected facets, the source or justification of testing and the function or outcome of testing." This broad framework has been criticised for being difficult to implement in practical terms. Kane (2007: 77) described this limitation as frustrating; that the "unitary framework may be more useful for thinking about fundamental issues in validity theory than it is for planning a validation effort". Sireci (2007: 478) echoed this view, saying that "the unitary conceptualisation of validity has done little to provide guidance regarding how to validate the use of tests in specific situations". How might this dilemma be resolved? About unified validity, Markus (1998: 77) accepted Messick's conjecture that some values are objective and others subjective, and that this idea partially resolves the tension between establishing validity from test evidence and validity as a consequence of values brought to the testing process. Markus (1998: 73) called this an approach based on value justification. This also seems to offer another, albeit problematic, avenue through which to explore the verification of validity. One cannot assume that all values are universally held in the same regard by all those involved in the test development process. Markus suggested that a value justification approach could allow for practical application of the unified validity theory to a validation effort on a case-by-case basis: "Both Messick and Moss (1998) emphasized the idea that the tensions involved in validity assessment are, to some extent, to be worked out case by case. That is, that the tension internal to the theory becomes a tension internal to the process of validation. This resonates with the previous point about some values remaining subjective and therefore relative inasmuch as that too implies that some things need to be worked out case by case" (Markus, 1998: 77).

2.4 Issues of Reliability

Whilst validity is a quality of the interpretation and the use of assessments, reliability is a quality of the assessment results themselves. Reliability estimates the consistency of a measurement, or more simply the degree to which an instrument measures the same way each time it is used under the same conditions with the same sort of subjects. "Methods of computing reliability therefore try to estimate the extent to which an observed score is near to a true score, for it is only if the observed score is a good estimate of a true score that we can draw sound inferences from the score to the construct" (Fulcher & Davidson, 2007: 105).

Porter stated that "without demonstrated reliability, there is no evidence other than anecdotal that a test is measuring anything, and its results cannot be trusted" (Porter in Alderson & North, 1991: 30), but Bachman (1990: 160) pointed out that in "any testing situation, there are likely to be several different sources of measurement error, so that the primary concern in examining the reliability of [assessment results] is first, to identify the procedures for estimating the effect of these sources of error on [the results]". In his introduction to his chapter on reliability, Bachman (1990: 163-4) identified several factors that may negatively affect the reliability of an assessment, which, as much as possible, should be considered when attempting to gauge the reliability of an assessment instrument:

- Factors that affect individuals' performance: "poor health, fatigue, lack of interest or motivation, and 'test-wiseness' (the amount and type of preparation or prior experience with a given assessment)"

- Factors that affect the assessment itself: "how a given test task is presented, the actual language use required by the task", and what Gipps and Murphy (1994: 19) refer to as "differential performance factors", or factors which are implicitly or explicitly biased.

As explained by various assessment experts (Weir, 2005; Hughes, 2003; Alderson, Clapham & Wall, 1995; Bachman, 1990; Reilly and Lewis, 1988), the reliability of a test can be measured in three different ways:

1. test-retest reliability: Test scores of the same test are about the same after 2 separate assessments to the same group.
2. equivalent (or parallel) forms reliability: Test scores are about the same for two different forms of the same test.
3. internal consistency reliability: Test scores are about the same for two different halves of the same test.

However, Alderson (1991: 65) interestingly stated that he believes "it is of little moment whether we term the concepts reliability or validity, provided that we address the issues that the labels conceptualise: legitimate and illegitimate variability and variation in the scores assigned to performance (whether from counts or ratings)". Alderson observed that what is usually considered and/or labeled *reliability* is in actuality, just another kind of validity. For example, he puts forward that both item homogeneity and item heterogeneity are not reliability issues, KR20 and Cronbach Alpha reliability coefficients notwithstanding, but rather a "matter of validity – whether the item or the test measures what it is supposed to measure rather than something else, whatever that something else is" (Alderson in Alderson & North, 1991: 62). And even when undesirable item heterogeneity is revealed by statistical analysis, we need to thoughtfully consider what this means: "Such heterogeneity only results in error in the sense that the test now measures more than one thing. It could still be doing so consistently, and therefore be reliable" (ibid.). In the same article, Alderson states also that test-retest reliability is indefensible (because of what he considers irresolvable issues associated with drawing useful conclusions from having participants sit for the same assessment twice), and that parallel form reliability is essentially concurrent validation. Weir (2005: 22) reiterated this conviction saying that all aspects of reliability may be termed "scoring validity" because there is a "growing consensus that [reliability issues are] a valuable part of a test's overall validity". The bottom line of this polemic may be that once the validity of a particular assessment has sufficiently been determined within the parameters of clearly defined uses for the tool, that reliability may reasonably be expected to exist.

2.5 Validity - The Discussion Continued...

Two other approaches to test validation will be considered here. Kane (2004) presented what he called ‘argument-based’ validation as a way of supplementing traditional construct validation. More recently, Lissitz and Samuelsen presented what has been called a radical simplification of “our conception of validity” (2007, in Kane, 2008: 76).

Kane: Argument-Based Validation

Kane (2004: 136), recognizing that the difficulty of applying validity theory to testing programmes is “exacerbated by the proliferation of many different kinds of validity evidence and by the lack of criteria for prioritizing different kinds of evidence”, put forth an argument-based approach to validity. It is a “methodology for evaluating the validity of proposed interpretations and uses of test scores” (Kane, 2004: 166)

“An argument-based framework for validation specifies the inferences and assumptions implicit in an interpretation or use of test scores and evaluates the plausibility of these inferences and assumptions (and of plausible alternative interpretations) using appropriate evidence. The evidence for and against the proposed interpretations and uses provides an overall evaluation of the validity of the claims based on the test scores. This approach tends to avoid the potential for reliance on largely irrelevant relationships that can occur in a weak program of construct validation” (Kane, 2008:79).

According to Kane (2006: 18), validation employs two kinds of arguments:

1. The development of an interpretive argument determines the proposed interpretations and uses of test results by identifying the inferences and assumptions.
2. The validity argument provides an evaluation of the interpretative argument which claims that a proposed interpretation is valid by asserting that the interpretive argument is coherent, the inferences are logical, and the assumptions are possible.

This is why Kane (2004: 166) says that this “argument-based approach is consistent with a unified model of validity based on the general principles of construct validity” and why this approach increases the strength of inferences made from a unified concept of construct validity.

One of the difficulties noted in the operationalisation of this approach is exactly how to do it, basically the same criticism Markus, Moss, Lissitz, Samuelsen and others have made about Messick's (1989) unified construct validity. "The devil is in the details. How does one craft a coherent interpretive argument? How does one systematically collect evidence to support that argument?" (Talbot and Briggs, 2007: 205) In defense of his argument-based validation approach, Kane (2004: 167) affirmed that while this "approach does not lead to a simple yes-no decision about validity, it does provide a way to gauge progress. In some cases , this process may uncover serious flaws that cannot be corrected…".

Lissitz and Samuelsen: Changing Terminology and Validity Emphasis

In their article, "A Suggested Change in Terminology and Emphasis Regarding Validity and Education", published in a special edition of the *Educational Researcher* journal, Lissitz and Samuelsen put forward a radically new vision of validity for the testing community to consider. As summarized by Kane (2008: 76), Lissitz and Samuelsen (2007) seem to be making four related points: "First, they argue against Messick's (1989) unitary view of validity based on construct validity… Second, they want to make an updated version of content validity the centrepiece of validity. Third, they claim that validity is a property of the test, independent of any proposed interpretation or use of the results. Fourth, they suggest a new vocabulary in which the updated version of content validity becomes essentially the whole of validity, and other issues traditionally housed under the heading of validity are to be considered separately as indicators of utility".

About Messick's unitary validity theory, they pointed out that its very complexity makes it less handy as a tool for test developers. "Messick's conceptualization and extensive discussion offers little in the way of useful advice to researchers interested in theory development or in the process of supporting or rejecting an established theory" (Lissitz and Samuelsen, 2007: 445). Sireci conferred with them about not only the inaccessibility of Messick's theory, but the central role content validation should play in test development: "It is far easier to talk about the content domain measured, particularly when it is operationally defined using test specifications. Thus the concept of content validity is much more palatable and understandable than the concept of construct validity to policy makers and the general public" (Sireci, 2007: 478).

Instead, they explained, "part of the problem with construct validity is that it is actually a compound concept, including both a construct-centred activity (what we call an internal focus) and a nomological activity (what we call an external focus)" (Lissitz and Samuelsen, 2007: 483). And they also proceeded to break it down further, citing "that the combination of theory and evidence regarding the constructs leads to a confounding that makes pursuit of validity (or theory) very difficult. It is the separation of these two intentions – construct definition and theory building – that we advocate and believe will improve the understanding and utility of both concepts" (Lissitz and Samuelsen, 2007: 440). They conceptualised this re-organisation in the following table, copied from the article:

Table 2.1: Taxonomy of test evaluation procedures (Lissitz and Samuelsen, 2007: 238)

		Perspective	
		Theoretical	Practical
Investigative Focus	Internal	Latent Process	Content and Reliability
	External	Nomological Network	Utility and Impact

Clarified in Table 2.1, "the approach that is being developed should initially be focused on the content elements of the assessment (what we call internal and theoretical). When this stage is completed, the researcher should move on to the external and theoretical" (Lissitz and Samuelsen, 2007: 440). So, according to these authors, in validating a test, the first consideration should be a careful consideration of the content: "It is the content validation process that ensures that the test reflects the domain of interest" (ibid.). In this way, they advise that we should first "study the measuring device (internal focus) and then study the nomological network (external focus, theoretical perspective)" (Lissitz and Samuelsen, 2007: 441). Sireci (2007: 480) also voiced his reservations about the minimization of the importance of content validity in Messick's theory: "The unitary conceptualisation of validity has undermined the importance of content validity in evaluating the utility and appropriateness of tests used in educational contexts" (Sireci, 2007: 480). Putting this idea of restructuring another way, they said, "One could argue that we have a choice: either expand and deepen our understanding of content validity to include the idea that we are constructing a construct, or change the definition of construct validity to minimize the role of the nomological network that deals with the development of the theory that relates multiple constructs to each other" (Lissitz and Samuelsen, 2007: 442).

Next, as Kane (2008: 77) pointed out, Lissitz and Samuelsen (2007) view validity as intrinsically a property of the test itself. "Messick's (1989) assertion that validity cannot reside in the test [is] essentially incorrect and confusing. We argue that it does, in fact, reside in the definition of the test, the development phase, and any psychometric theory that gave rise to that test and its associated constructs, whether latent or manifest" (Lissitz and Samuelsen, 2007: 442). They added that they "are suggesting that lack of empirical evidence in support of a theory is not an indicator of inadequate content validity evidence for the test device. Our notion of validity clarifies the confusion about whether the theory or the rest is at fault when construct validity is not found. In our system, the fault is in the theory because theory development would always be the result of a two-stage process in which determination of the content validity sense of the measure precedes the use of the test in a theory. This approach is consistent with that of Holland (1988, in Lissitz and Samuelsen, 2007: 445), [who said] what the first thing to do is to determine the measure and only after that is done can one move to building a theory involving that measure".

Lissitz and Samuelsen argue therefore, that content validation should be pre-eminent, and that other types of validation which draw on external (outside of the test itself) variables to support validations should be seen as important, but secondary. "The internal characteristics of an instrument do not depend on the relationship to some external variable to define the instrument's validity, although such external variables can be important. The other characteristics – what are currently called criterion and construct validity determinations – are very important, but the user of these techniques should, we argue, recognise that they answer fundamentally different questions. It is for this reason that they should not be presented as a unified theory of validity" (Lissitz and Samuelsen, 2007: 446). In addition, they state that the impact of a test upon stakeholders is not a validity issue at all. "Messick (1989) referred to the validity at issue in such studies as consequential validity. … If a test is shown to have some impact that is unintended or unwanted, that observation should not be considered relevant to the question of whether the test is valid. Impact evaluation is not a characteristic that helps determine content validity, but it can be an important consideration" (Lissitz and Samuelsen, 2007: 445).

However, they regard the issue of test score reliability as an internal test variable, along with content validity: "We suggest that these essentially internal characteristics (reliability and content validity) be called the *internal validity* of the test, and all other characteristics be considered essentially external matters" (Lissitz and Samuelsen, 2007: 446). This is because they view reliability as independent of external measures of validation. Reliability "statistics are essentially a characteristic of the test and do not depend on any empirical or theory-driven study of relationships with other measures external to that test. Reliability is also a fundamental characteristic of a test that does not depend on other external measures for its meaning" (Lissitz and Samuelsen, 2007: 446). This is also clearly represented in their taxonomy in Table 2.1. Having made this declaration, it seems almost contradictory for the authors to state in their 'response to comments' article that the "concept of reliability is not a completely different concept, as often presented, but shares many characteristics with validity, particularly in the area that we term utility, traditionally called criterion validity" (Lissitz and Samuelsen, 2007: 482). The authors did speak of variables overlapping across the lines of their taxonomy, and perhaps this is one instance of such overlapping. In support of this possibility, the authors also speak of reliability as a 'critical descriptor'. "In this new formulation of validity, the test definition and development process (what is currently known as *content validity*) and test stability (what is currently known as *reliability*, or sometimes *generalizability*) become the critical descriptors of the test. They also become the primary justification for its existence and acceptance for use. They exist independent of, or regardless of, the application of the test or the use of the test in some theoretical formulation" (Lissitz and Samuelsen, 2007: 446, italics in the original).

The criticism of Lissitz's and Samuelsen's validity re-think which enjoyed the most unanimity amongst the commentators was their belief that construct verification should not play an important role in test development. "The class of theory support studies, again, has no direct impact on the internal validity of an assessment device. … We would say they are engaging in work to support a theory. The failure to verify a theory involving a particular construct as indicated by a particular test and choice of psychometric theory should have, obviously, no impact on the belief in the validity of that test as a measure of some construct" (Lissitz and Samuelsen, 2007: 444). And even more strongly stated in the follow-up article: "We have taken the position that theory is important but has gained little or no benefit from the 50 or more years of ruminating about construct validity" (ibid.). The only theory which they see intimately involved in test construction is psychometric theory (Lissitz and Samuelsen, 2007: 446).

In their comments about the original article, Both Sireci (2007: 480) and Moss (2007) felt that the proposed changes may be "relevant mainly to educational tests. In other testing contexts, such as personality assessment and projective testing, certainly construct theory must take a more prominent role". Another important reason Sireci (ibid.) pointed out was that "theory is needed to design validation studies and to rule out the plausible rival hypotheses that may explain test performance".

Moss (2007: 479) noted that "a theory of validity represents both a philosophical perspective and a set of conceptual tools that shape our thinking and action. … The separations that Lissitz and Samuelsen propose for the way we think about validity – between the theoretical and the practical, the internal and the external, the purpose of testing and the development of the test – risk losing the power of the principles of scientific inquiry to guide the evaluation…". Kane expressed concern about the resulting lack of depth of test validity if such changes were implemented: "Their conceptualisation would certainly make validation more manageable, if not fairly easy; evaluate the representativeness of the content and check on reliability and scaling, and the test would be validated. However, the interpretation that would be validated would be a narrow, operational interpretation, and no attempt would be made to justify any use of test scores" (Kane, 2007: 76). Sireci (2007: 481) concurred with the others saying, "Validity evidence based on test content is necessary but not sufficient. A serious effort to validate use of an educational test should involve both subjective analysis of test content and empirical analysis of test score and item response data". Moss was the strongest of the commentators in her support of Messick's unitary validity theory: "A unitary conception of validity is in no way inconsistent with the provision of substantial guidance, nor does it preclude the making of well-reasoned, practical judgements about what can and should be undertaken before and after a test is put into operational use" (Moss, 2007: 470). Messick's Unitary Validity Theory has received wide support, but serious operational impediments remain, like the establishment of construct validity. To sum up, validity has long been held to be a valuable and necessary test quality, but the discussion amongst scholars about how best to ascertain validity – in all its forms – continues.

2.6 Pesky Predictive Validity

Predictive validity may be considered 'pesky' for its perseverance amongst various test stakeholders as a test quality which is desirable, despite the weaknesses that professionals in the field of assessment criticise it as possessing. There have been an abundance of studies since 2000 which have investigated the predictive validity of different testing instruments with regard to language competence and/or academic achievement. Some examples of such studies are: Ehlert and Podgursky's predictive validity study of a high school state assessment (2005); Ortega and Payne's predictive validity of the GRE (2007); Kuncel and Hezlett's study of the predictive validity of 7 different graduate gate-keeping exams (2007); Wei's predictive validity study of preliminary examinations and O level results (2006); Paul's study of the IELTS as a predictor of academic language performance (2007); Ingram and Bayliss' study of the IELTS as a predictor of academic language performance (2007); Sternberg's study of the predictive validity of the American SAT (2006); Tanner's study of the predictive validity of 8 different placement exams, 2003); Geiser and Studley'sb predictive validity study of the American SAT (2002); Fleming's predictive validity study of the American SAT with African Americans (2000); Bridgeman, McCamley and Ervin's predictive validity of the revised American SAT (2000); Armstrong's predictive validity study of a placement exam (2000); Stofflet, Fenton and Straugh's predictive validity study of a high school state assessment (2001); and others. (The year 2000 was a fairly arbitrary cutoff year for the researcher with regards to predictive validity study inclusion in this review, owing to the widely-held consensus that in academic research, it's generally preferable to cite references on more recent comparable data.)

Reaching back to 1989 for a definition from which to proceed, Harrison (1989: 145) explained that "predictive validity has to do with the predictive force of a test." Theoretically, the test is valid if it is able to reasonably predict future academic performance. In this kind of validity, we take a sample of a student's performance in a test and generalise that sample to the universe of his understanding in order to assess what we hope he has learned (Nuttall, 1986: 2). The predictive validity of an assessment may also aid teachers in choosing what they need to concentrate upon developing with the students in the future, and what areas of learning each particular student needs to concentrate upon in the future. Messick (1989: 41) observed that "with respect to the generality of the process, the development of evidence to support an inferential leap from an observed consistency to a construct or theory that accounts for that consistency is a generic concern of all science". Bachman (1990: 254-255) restates this in another way by asserting that "construct validity concerns the extent to which performance is consistent with predictions that we make on the basis of a theory of abilities, or constructs".

Notwithstanding its popularity, predictive validity has come under serious criticism. Brown (1990: 42) pointed out that "prediction of future performance is always dubious. There is no certainty that motivation and opportunity to achieve will remain constant over time." Therefore, "assessments and records of what students have achieved cannot and should not be treated as predictive of what students can and will achieve [in the future]" (Sutton, 1991). McNamara (2001: 337) also stated that "our existing models of performance are inadequately articulated, and the relationship between performance and competence in language testing remains obscure. In particular, the assumption of performance as a direct outcome of competence is problematic, as it ignores the complex social construction of test performance". As these authors point out, not only is the grading function being over-emphasised and the learning function under-emphasised, the whole social aspect of performance in testing is not well-understood or appreciated. However, formal, summative examinations dominate students' lives, since it is the scores they receive from these examinations that will determine the opportunities they will have for higher education and future employment (Broadfoot, 1984: 123).

These problems are exacerbated when second language students are involved. Short (1991: 2) explained the two-edged sword of standardised, external examinations upon students whose English ability is described as "limited". As she explained, we need to be very concerned about what we are actually testing because "teachers may not be sure whether a student is simply unable to demonstrate knowledge because of a language barrier or whether, indeed, the student does not know the ... material being assessed. [B]ecause language and content are intricately intertwined, it is difficult to isolate one feature from the other in the assessment process." This criticism of the use of standardised tests as selection criteria for ESL students was echoed by Brenda Denvir (1989, in Rumsey, 1998: 10) when she said that "they have a long-lasting and negative influence on those not selected for whatever benefits are offered".

Acknowledging these criticisms, however, does not negate the fact that these results have been and will continue to be used for selection criteria (aka gatekeeping). As Paul Black said (1998: 44), "It would be easy if all those who took a test were admitted to degree study, irrespective of their results, but this does not happen". Absolutely. This does not happen for two reasons: 1) It's obviously not possible for institutions to admit everyone who applies. There is and will always be a need for some assessment instrument to use as part of the selection criteria; and 2) Admitting everyone to a degree programme, regardless of their results, would also obviously negate the need for the assessment instrument to begin with. It is imperative that assessment experts come to terms with the reality of this situation while recognizing the limitations of this type of validity reference. We need to be very cautious about the predictions we make on the basis of students' scores, but we cannot allow ourselves the intellectual luxury of saying that this type of validity is not defendable because of the truncated sample or the effect of a myriad of other variables, some known and some not. As we have seen, the strength of any type of validity regarding the conclusions we come to is distorted by what Messick (1998: 37) referred to as the "consequential basis of test validity".

Differential validity and differential prediction refer to two observed measurement phenomena. Differential validity refers to group differences in correlations between a predictor and a criterion (validity coefficients). Differential prediction refers to group differences in the best prediction equation and/or the standard errors of estimate (Young, 2001: 290). The presence of differential validity and/or differential prediction may indicate lack of fairness in test use. Differential prediction is considered the more serious problem of the two because differences in prediction have more direct impact on fairness in selection (Linn, 1982: 281). For example, in the context of college admissions, differential prediction implies group differences in the prediction of an outcome, such as college GPA. Using multiple regression analysis and categorical indicator variables, significance tests may be conducted (for equality of the regression equations) to determine the degree of differential prediction and validity.

2.7 Referencing

Another area of general agreement in the area of assessment terminology refers to the reason for using a particular assessment vehicle. The CEPA exam is a norm-referenced examination. In a norm-referenced test, as described by Bachman (1990: 72) and Bailey (1998: 35), the quality of a student's performance is judged by comparing it to other students in the same group. Norms are not static, but constantly changing as a result of societal or political change or a shift in paradigmatic focus (Broadfoot, 1984: 24). As Alderson, Clapham and Wall (1995: 76) noted, "In many examination systems it is the case that objectively marked tests are treated as norm-referenced, and subjectively marked ones as criterion-referenced. This is probably not from any underlying testing philosophy, but because of practical considerations". Bachman (1990: 74) further described standardised tests as the "quintessential" norm referenced test, because they all have 3 characteristics: (1) They're based on a "fixed or standard content, which does not vary from one form of the test to another", and alternate forms of the test "are carefully examined for content equivalence"; (2) "there are standard procedures for administering and scoring the test, which do not vary from one administration of the test to the next"; and (3) the test has undergone a rigorous process of empirical research, development and trialing. The National Admissions and Placement Office (NAPO: See page 5 for more information.) has demonstrated that the CEPA possesses these characteristics to a great degree through published research (i.e. Brown and Jaquith, in O'Sullivan, 2011) and in conference presentations, as well as their website (http://ws2.mohesr.ae/napo/Details_EN.aspx?str=SR last accessed July, 2012).

An important point to consider is that some of these international examinations are consciously produced, and therefore norm-referenced for, use in a particular country, thereby putting at a disadvantage those sitting for the examination outside the country (Gibbs and Murphy, 1994). Table 2.2 (adapted from Bachman, 2011: from course outline, "Foundations of Language Assessment" at UCLA) outlines other disadvantages, as well as advantages of norm- and criterion-referenced exams.

Table 2.2: Advantages and Disadvantages of Norm-Referenced and Criterion-Referenced Exams

	Advantages	**Disadvantages**
Proficiency (NR)	1. Can maintain desired numbers of students in different course levels by choosing appropriate cut percentiles 2. If cohorts vary in size from year to year, can maintain desired numbers of sections in different course levels by choosing appropriate cut percentiles 3. If course syllabi change, no need to change test	1. If cohorts vary in proficiency from year to year, there will be shifts in proficiency levels for various course levels, perhaps resulting in inappropriate placements 2. Need to develop separate test batteries for placement, progress and grading decisions
Achievement (CR)	1. Potential for more appropriate placements, with respect to course content objectives 2. Potential for providing teachers with diagnostic information related to course objectives 3. Only need to develop a single test battery for placement, progress and grading decisions	1. If cohorts vary in proficiency or size from year to year, there will be shifts in the numbers of students in various course levels, but placements will still be appropriate 2. If course syllabi change, need to change test

2.8 MCQs

The CEPA examination, which is the focus of this study, is entirely composed of multiple choice questions (MCQs), except for a short essay section which is banded separately. The objective part of the CEPA covers reading comprehension, grammar and vocabulary. It is important, at this point, to consider the rationale for using this format in general, as well as its pros and cons. One possible rationale for choosing MCQ format is the status quo: most international, standardised English proficiency examinations use this format. "From the mid-1960s, through the 1970s, language testing practice was informed essentially by a theoretical view of language ability as consisting of skills (listening, speaking, reading and writing) and components (e.g. grammar, vocabulary, pronunciation) and an approach to test design that focused on testing isolated 'discrete points' of language, while the primary concern was with psychometric reliability (e.g. Lado 1961; Carroll, 1968)" (Bachman, 2000: 2). Even though this notion of testing discrete points of language has faced criticism from language testing experts, it has persevered. As Porter (in Alderson & North, 1991: 30) stated, "Although it is by no means universal, the notion of the 'four skills' does indeed show a remarkable resilience among testers, teachers, examining boards and consumers of test results in general. However, the fact that many people find it convenient to describe language ability in terms of four separate skills, while many others would wish to describe it in other ways, involving integrated skills, suggests that these two modes of description and testing simply capture different truths; the superiority of either over the other has yet to be demonstrated" (Porter in Alderson & North, 1991: 30).

This test format enjoys strong face validity amongst those not involved in assessment professionally: "...multiple-choice testing meets many of the objections about the inaccuracies of examinations and may command a considerable degree of public acceptability, especially since its apparently objective, systematic techniques fit well with the prevailing technocratic ideology" (Broadfoot, P., in Nuttall, D, 1986: 59). It has been said that this sort of "…conservatism in testing exists because there is a tendency for old solutions to be relied on too much, leading to conservatism in test formats; there are not enough theory-driven pressures for change; and there are not too many separate groups of testers with different goals which do not complement each other" (Skehan, in Alderson and North, 1991: 6).

About the MCQ test item format, Purpura (2004: 129) pointed out that a multiple choice question requires the test taker "to choose the correct answer from the response options given". There are usually 3, 4, or 5 options, and obviously there should be only one best answer. In test jargon, the answer is called the item key, and the other options are called distractors. (The test question itself, which is not necessarily always a 'question', is called the stem). This is the format of the CEPA.

Hughes (2003: 76) wrote about the benefits of MCQs: "Perhaps the most obvious advantage of multiple choice is that scoring can be perfectly reliable. Scoring should also be rapid and economical. A further considerable advantage is that, since in order to respond the candidate has only to make a mark on the paper, it is possible to include more items than would otherwise be possible in a given period of time. ... Finally, it allows the testing of receptive skills without requiring the test taker to produce written or spoken language." The so-called benefits of objective scoring and ease of administration were also echoed by Purpura (2004: 132), but he also mentioned the additional advantage of the pre-testing of the questions, "so that their psychometric characteristics can be determined prior to operational testing. In this way, 'easy' or 'difficult' items can be selected and ordered as needed".

Along with the advantages to be had with this test format, there are serious disadvantages in both the development of MCQ items and the effect this format has on the test takers themselves, including test-taking behaviour/s. "There is evidence that students taking multiple-choice tests can learn strategies for taking such tests that 'artificially' inflate their scores: techniques for guessing the correct answer, for eliminating implausible distractors, for avoiding two options that are similar in meaning, for selecting an option that is notably longer than the other distractors, and so on. There is also evidence from anecdotal accounts of multiple-choice test takers that the test method tends to encourage students to consider alternatives they would not otherwise have considered..." (Alderson, Clapham & Wall, 1995: 45). To sum up, Broadfoot (2005: 129) asked us to consider even the assumed "objectivity" of this format: "Even the most apparently objective assessment – a multiple-choice test – is objective only in its scoring: it is not an objective assessment as such simply because all assessment involves professional judgement". Purpura (2004: 132, and others, i.e. Bachman, 1990: 129) have pointed out that the development of well-written MCQ items is difficult and time consuming. It is often said that items that are easy to write are difficult to mark and vice versa.

2.9 Scaled Scores

There are two types of test scores: raw scores and scaled scores. A raw score is a score without any sort of adjustment or transformation, such as the simple number of questions answered correctly. A scaled score is the result of some transformation applied to the raw score. As explained by the CEPA supervisor (See page 146.), the score for the CEPA is a scaled score.

The purpose of scaled scores is to report scores for all examinees on a consistent scale. This can pose difficulties if a test has two (or more) forms because it may be that one is more difficult than the other. This situation can be somewhat resolved by piloting the forms, which is what NAPO does for the forms of CEPA. Scaling is something that occurs after the assessment process is completed (including equating), what Weir (2005: 22) termed "a posteriori validity evidence". The idea of scores being a measure of validity rather than reliability will be discussed further on. (Comparing the difficulty of a test form to another is called making equivalent forms or "equating". For example, an analysis may determine that a score of 65% on form 1 of a test is equivalent to a score of 68% on form 2. Scores on both forms can be converted to a scale so that these two equivalent scores have the same reported scores.) Two well-known tests in the United States that have scaled scores are the ACT and the SAT. (http://www.docstoc.com/docs/2259128/Understanding-a-Scaled-Score-What-is-a-scale; accessed 15/7/2011)

A reporting scale that remains constant across different forms of the test enables comparisons to be made. The scale adjusts for the difference in difficulty to provide the same standard score for the same level of performance. Each time a test is given, different questions in different combinations are presented. "These procedures help to ensure comparability of difficulty and maintenance of standards across different forms of the [test] and between sessions." (Weir, 2005: 27) This helps to ensure that the exam is fair to candidates. Because each test session is made up of different combinations of questions, the test a candidate takes the first time is different from the test he or she takes the next time. Although great care may be taken to ensure that each form is perfectly parallel in content and difficulty, there may be variations from one form to the next. The careful selection of questions from item banks ensures different forms sample the same content areas. The CEPA has a large item bank, as do many international standardized exams. Weir (ibid.) explained that the process used by Cambridge ESOL exam papers included the construction of an underlying scale "onto which the difficulty of all the test items [in the item bank] can be mapped across the five examination levels". He further explained that the scale is "achieved by routine pre-testing of new items alongside items or test components with known difficulty values (i.e. anchor tests). All new items are now calibrated in this way using Rasch analysis, and are put into the item bank with values linking them to the common scale" (Weir, 2005: 27).

Alderson countered that "there is a tendency to treat test scores as being on an equal interval scale; this is simply not the case. Items vary in their difficulty, complexity and demands made on learners' proficiency" (Alderson in Alderson & North, 1991: 85). In spite of the fact that care is taken by developers of large-scale, high-stakes standardised proficiency exams to ensure that the scores test takers receive are as true a reflection of the achievement of a certain competency level as possible, caution has been voiced by experts about becoming over-confident with regards to this. Alderson (1991: 65) pointed out that "[i]n all educational measurement, what we hope for and work towards is that the score reported reflects the candidate's ability". He, with North (Alderson & North, 1991: 84), further elaborated upon this sentiment, saying that there are recognized difficulties with the "issues associated with producing scales of proficiency and their related band scores and descriptions, [but] the same problems exist with test scores. … How one interprets a score of 30/50 on a given test will depend upon factors like: the difficulty of the test; the nature of the test's content and the method used to assess mastery of that content; the performance of other students; the performance of 'criterion' students… It is [also] clearly the case that different individuals who achieve a score of 30/50 will have got different items right and wrong. Thus, no one score of 30 will be equivalent to other scores of 30, and the same scores will have been achieved by individuals with different constellations of ability. … The comparability of scores becomes a real problem. … There is no immediately obvious practical solution, since it is clearly impossible to report candidates' performance on each individual item, rather than on the test as a whole". It would seem that this is an irresolvable issue inasmuch as testing is a ubiquitous element of assessment in general and admissions criteria in specific, and the apparent impossibility of determining with reliable precision score interpretations about test-takers' responses to various test items. This ‘irresolvability’ also strengthens the case for the necessity of basing admissions decisions on several different criteria.

2.10 Intra-Rater, Inter-Rater Reliability and Band Scales

As clearly defined by Weir (2005: 34), in marking tests of writing, the most important consideration is the consistency of the people who mark them. Weir (2005: 34) went on to explain that “[m]arkers need to be consistent in two ways: each marker needs to be consistent within himself (intra-rater reliability), ... and there needs to be consistency of marking between markers (inter-rater reliability). In many tests, two raters are used”.

There seems to be some variance of opinion with regards to which marker is more effective and astute: human raters or computer rating programmes. Some have stated that human raters cannot approach the consistency of computers. Lim (2009: 57) mentioned that “certain test providers have begun introducing automated rating by computers (e.g., the use of e-Rater by the Educational Testing Service)”. O’Sullivan (in Simpson, 2011: 268) stated that “automated scoring systems allow for what is known as ‘person-free’ assessment of the written performance of learners and test-takers. Given the subjective and idiosyncratic of human ratings (see the work of Lumley (2005) and O’Sullivan (2008) who explored the nature of rating in tests of writing and the impact of interlocutor-related variables in testing speaking respectively), it is difficult to ignore the claims made by advocates of these automated systems, particularly when they report very similar outcomes to those of human raters, but with much higher consistency (see, for example, Foltz *et al*. 2000)”. Having said this, it must be noted that the effectiveness of these systems does depend on having a large enough database of essay writings (on similar topics), marked by trained, human raters, which this software can use to identify features typical of work at particular skill levels.

However, trained, qualified raters are considered by many a vital part of the assessment process for writing exams. Just like the automated systems, the human rater’s role is an intermediary one between the test takers and the final marks awarded. Writing exams are subjective and can include discoursal and stylistic elements that are quintessentially human, and problematic for computers to digest. Human raters, technological advancements and innovations notwithstanding, do continue to be primarily relied upon for writing exam scoring (Lim, 2009: 57). Despite this, discrepancies in rater consistency may negatively affect the validity and reliability of writing exam results. In his investigation of variability in the IELTS General Training Writing Module, O'Sullivan (2002: 14) stated that "[o]ne measure of the value of a test is the degree to which construct-irrelevant variance intrudes on reported performance. This variance constitutes a serious threat to the validity of any inferences we wish to draw from test performance, in that it means that the test score is not wholly representative of the ability being measured, but is an amalgam of variance that can be attributed to the ability being tested as well as any 'noise' caused by outside factors". For example, an examiner or test administrator who is exhausted, bored, or distracted may incur a negative effect upon the writing that he/she is assessing. "In order to account for features affecting estimates of marker reliability in speaking and writing tests, it is now common to use sophisticated IRT models. For example, Multifaceted Rasch (MFR) analysis may be applied. MFR has clear importance for detecting inconsistent individual marker behaviour both over time and in comparison with other markers." (Weir, 2005: 35) This is because the most commonly-used Rasch analysis (the Dichotomous Rasch) considers "the measurement problem with an object of measurement (e.g. student, applicant) with an agent of measurement (e.g. test, questionnaire)" (Mead, 2008, 23). It is a fairly straightforward operation to use the key to correct the assessment and assign a score. However, "many situations are not so mechanical and there is some judgment involved in assigning the score. No matter how well-trained or experienced, judges will sometimes differ" (Mead, 2008: 23). MFR can take their differences into account, dichotomous Rasch cannot.

Different strategies have been developed and implemented over the years to try to address and reduce human scoring inconsistencies. Bailey (1998: 185-203) outlined three approaches to scoring writing exams: holistic, analytic and objective. As she explained, "[r]aters using holistic scales are trained not to think about the individual components of the writing skill ... [but] react to the student's composition as a whole". The advantages of this method, she added, are that it is "fast, high rater reliability can be achieved, and the scoring scale can provide a public standard understood by [all]. Holistic scoring systems are also assumed to be widely applicable to many different topics. In addition, many people feel that holistic scoring emphasizes the writers' strengths, rather than their weaknesses" (Bailey, 1998: 189). The training of raters involves familiarising them with "benchmark" papers that are good examples of each level (or scale/band). One important area in which improvements have been made is in rating scales (aka band scales and rubrics). This has greatly enhanced rater reliability, as Bailey (1998: 77) pointed out, "Written level descriptors and highly codified procedures of rater training have been developed to insure reliability in rating ...responses in ... tests of writing". This was echoed by Purpura (2004: 120), who pointed out that "[s]everal testers (e.g., Alderson, Clapham and Wall, 1995; Weigle, 2002) have cited ways of minimizing sources of bias and unreliability. Some of these ways include (1) using a clear and detailed scoring rubric; (2) training the raters; (3) using samples of performance in the rater training session that exemplify the different points on the rubric; (4) scoring performance independently by two raters, with a third to adjudicate in the case of large discrepancies between raters; and (5) monitoring rater performance and providing raters with constructive feedback". Some advantages of reporting scores on [band] scales instead of as raw or transformed test scores were outlined by Alderson (Alderson in Alderson & North, 1991: 85):

- "Scales can provide information about a learner's behaviour, and thus help a user to understand what a score means.
- Scales can help reduce the spurious impression of accuracy that a score gives.
- Scales can help increase the reliability of subjectively judged ratings, and provide a common standard and meaning for such judgements.
- Scales can provide guidance to test constructors wishing to write tests."

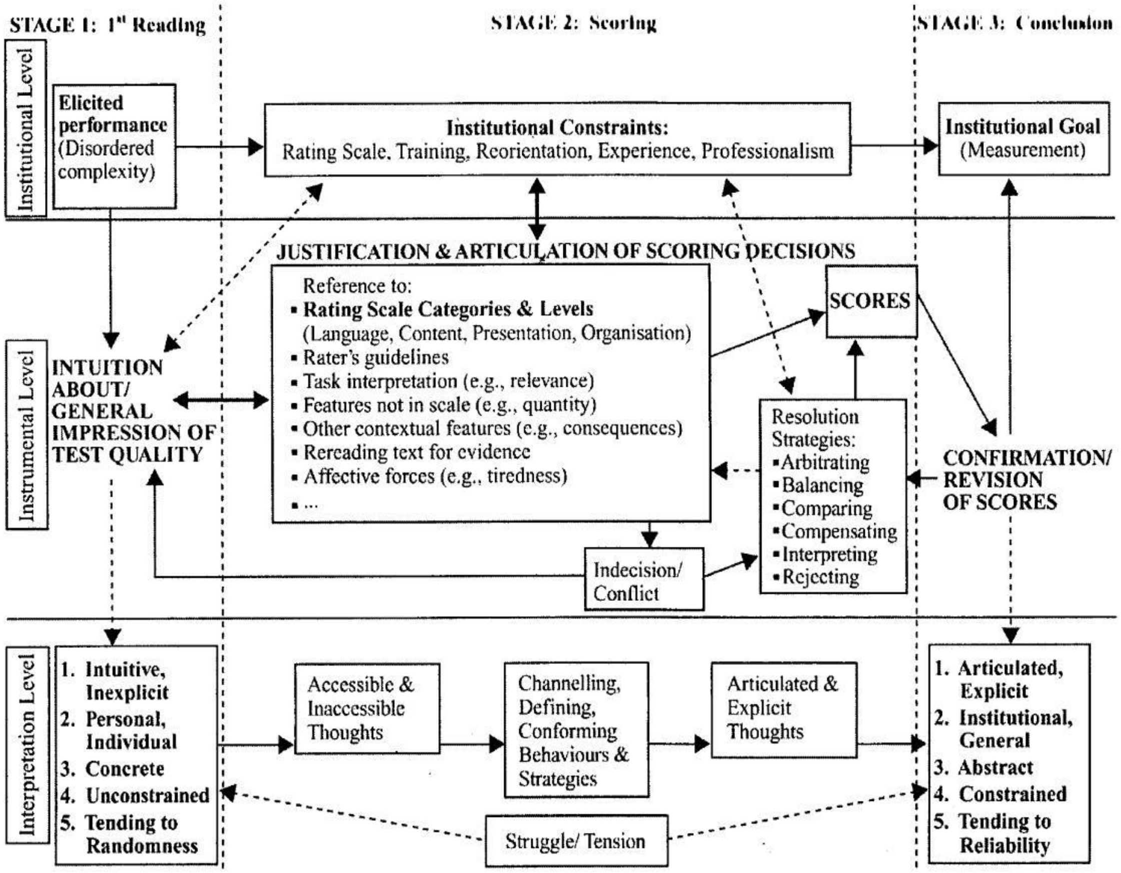
According to Bailey (1998: 190), analytic scoring also involves the use of scales to mark writing samples, but the components of the scale are weighted differently to reflect theoretical views of the importance of different input. For example, at HCT, "communicative competence" of the writing sample is weighted double the other components (mechanics, spelling, vocabulary, etc.) because the focus in that institution is upon how well a student *communicates* through his writing. The training of raters for analytic scoring goes through a similar, but perhaps more exhaustive, period of examining benchmark papers and norming the raters' scores than holistic raters. Barkaoui (2010: 54) studied the use of scales and raters' behaviour while using them by employing "think-aloud protocols to examine the roles of rating scales, rater experience, and interactions between them in variability in raters' decision-making processes and the aspects of writing they attend to when reading and rating ESL essays". He discovered that "with holistic scoring, raters tended to refer more often to the essay (the focus of the assessment), whereas with analytic scoring they tended to refer to the rating scale (the source of evaluation criteria) more frequently; analytic scoring drew raters' attention to all evaluation criteria in the rating scale, and novices were influenced by variation in rating scales more than were the experienced raters" (Barkaoui, 2010: 54).

Perhaps a happy medium in scoring writing can be established through a combination of technology and human participation. O'Sullivan (2011: 269) pointed out that "major examination boards have been exploring the feasibility" of online script marking, and he identified one experiment in particular – the one devised by CEPA for scoring writing: "In operation now for several years, the CEPA system is an excellent example of what can be achieved with the intelligent use of technology (see Brown and Jaquith, 2011)." In the correction of scripts from the CEPA, not only are the original scripts scanned and made available to raters online, the scripts are double-marked, and the raters' scores are also monitored by software that automatically corrects for excessive variance in a particular rater's marking.

However, as Lim (2009: 57) pointed out, "the desirability of increasing agreement by and of itself has come under question (Connor-Linton, 1995a; Lumley & McNamara, 1995; Reed & Cohen, 2001; Weigle, 1998). The inter-rater reliability statistic only says something about the product of assessment but not about the process. ... Raters could well be agreeing on things that have nothing to do with what is being measured. Thus, there is the need to better understand the rating process itself, as well as the rater characteristics that could affect raters‵ rating behaviour". At the conclusion of his study, Lumley (2006) offers what is perhaps the most complete model of the rating process to date (Figure 2.1, following page). There are three basic stages to the rating process: reading and prescoring, scoring, and revising and finalizing of scores. It can also be seen that this process takes place on three different levels, what Lumley calls the institutional level, the instrumental level, and the interpretation level.

The instrumental level, showing rater behaviour, is the most visible part of the process and, perhaps for that reason, what most studies have focused on. At the beginning of the process, raters read compositions and develop an intuitive impression of their quality. This stage is important because even though a rating has not been awarded, a judgment about texts has been made. For this reason, this stage is also called the prescoring stage. As to the source of those impressions, studies suggest that raters‵ backgrounds play a part (Cumming, 1990; Cumming, et al., 2002; Pula & Huot, 1993).

Figure 2.1: Lumley's (2006: 291) Model of the Rating Process



STAGE 1: 1st Reading
STAGE 2: Scoring
STAGE 3: Conclusion
Institutional Level
Elicited performance (Disordered complexity)
Institutional Constraints: Rating Scale, Training, Reorientation, Experience, Professionalism
Institutional Goal (Measurement)
JUSTIFICATION & ARTICULATION OF SCORING DECISIONS
Instrumental Level
INTUITION ABOUT/ GENERAL IMPRESSION OF TEST QUALITY
Reference to:
▪ Rating Scale Categories & Levels (Language, Content, Presentation, Organisation)
▪ Rater's guidelines
▪ Task interpretation (e.g., relevance)
▪ Features not in scale (e.g., quantity)
▪ Other contextual features (e.g., consequences)
▪ Rereading text for evidence
▪ Affective forces (e.g., tiredness)
▪ ...
SCORES
Resolution Strategies: ▪Arbitrating ▪Balancing ▪Comparing ▪Compensating ▪Interpreting ▪Rejecting
CONFIRMATION/ REVISION OF SCORES
Indecision/ Conflict
Interpretation Level
1. Intuitive, Inexplicit 2. Personal, Individual 3. Concrete 4. Unconstrained 5. Tending to Randomness
Accessible & Inaccessible Thoughts
Channelling, Defining, Conforming Behaviours & Strategies
Articulated & Explicit Thoughts
1. Articulated, Explicit 2. Institutional, General 3. Abstract 4. Constrained 5. Tending to Reliability
Struggle/ Tension

It is perhaps worth noting, to conclude this section about the use of rating scales to mark writing samples, the point made by Weir (2005: 191): "Choice of rating scale, as always, depends on the prevailing situation in the context of administration. In those cases where large numbers of scripts have to be marked fairly quickly... practicality might lead us to using a holistic scale because of its speed and ease of application." This leads us to the next section.

2.11 Issues of Practicality

Another criterion for determining the worthiness of a particular kind of assessment has alternately been identified as "usability" (Lyman, 1991), "manageability" (Sutton, 1991), "feasibility" (Alderson, et al., 2005; Nevo and Shohamy in Alderson, et al., 2005) or "practicality" (Gronlund, 1976; Weir, 2005). It refers to factors outside the test or exam itself that have very real implications for its development, administration and marking, like the limitations of time and cost. For example, multiple choice tests are sometimes appear to be more practical than other exam types because they can be machine-scored and the "preset scoring criterion is *right* or *wrong*, so no judgment is involved on the part of the scorer" (Bailey, 1998: 77; italics in the original, underscore mine). Acknowledging the importance of practicality, Weir (2005: 49) strongly cautioned against prioritising it, even leaving it out of his model for the establishment of validity, reasoning that "[o]nly when sufficient validity evidence is available to justify interpreting test scores as an acceptable indication of the control of an underlying construct, should we concern ourselves with practicality". He went on to explain that "[p]racticality considerations are often allowed to intrude at too early a stage and validity is often threatened rather than enhanced as a consequence" (Weir, 2005: 49). This point of view echoes Hughes (2003: 56) in his discussion of beneficial washback, in which Hughes also noted the importance of test practicality, saying: "...it is good that a test should be easy and cheap to construct, administer, score and interpret. We should not forget that testing costs time and money that could be put to alternative uses". However, like Weir, he countered by pointing out the real potential cost of not doing it right: "When we compare the cost of the test with the waste of effort and time on the part of teachers and students in activities quite inappropriate to their true learning goals (and in some circumstances, with the potential loss to the national economy of not having more people competent in foreign languages), we are likely to decide that we cannot afford not to introduce a test with a powerful beneficial backwash effect" (Hughes, 2003: 56).

But is this point of view 'practical'? Even though it was written much earlier, an observation of Davies' (1990: 6) seems to almost be in response to such assertions. He argued that if a test is not practical (because "it would take up too much time, too much skilled manpower, or it might require expensive or elaborate media systems or scoring arrangements", etc.), then "it must lack validity because it is unusable on its target population. And yet it is possible to establish reliability and validity on a laboratory sample and then, as a result of that application, to recognise the impracticality under less favourable conditions".

The difficulty of the consideration of practicality was acknowledged by Fulcher and Davidson (2007: 128-129) in the term 'constraints'. They defined 'constraints' as being "limitations on any of the resources that are necessary for successful test administration", and they listed seven areas in which these limitations could occur: in people, skills, equipment, accommodation, security, information technology, and money. Time constraints are another issue, and include how long it takes to administer the test and how long it takes to amass and calculate the results.

There are cost constraints, and these include not only the financial cost of the assessment itself as well as the cost of the professional(s) who administers the assessment and who determines the results, but also the cost effectiveness of the test system. The consideration of cost was included in Nevo and Shohamy's (1986, in Alderson, et al., 2005: 250) article in which they proposed a framework of Standards for Educational Testing Methods. They divided this framework into four main areas. The third area of standards in this framework was "Feasibility Standards". As quoted by Alderson, et al.(2005: 250), these standards "are intended 'to ensure that a testing method will be realistic, prudent and frugal'. The issues [they presented were] Practical Procedures, Political Viability and Cost Effectiveness".

In a different interpretation of the term 'practicality', Bachman and Palmer (1996: 18) named it as one of six characteristics of the 'usefulness' of a test in the form of an equation: "Usefulness = Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality". As Fulcher and Davidson (2007: 15) pointed out, "Bachman and Palmer used the term 'usefulness' as a superordinate in place of construct validity, to include [these test characteristics]". As interesting as this polemic may be, it has not 'caught on' in the testing community as a way of describing test usefulness. As with many other issues in test development, the struggle between what could be or would be ideal and what is feasible and practical is an important and often difficult issue to negotiate.

2.12 Testing Ethics: the concept of 'Fairness'

In 1997, Shohamy (1997: 341) noted that, "...research in language testing has always focused on [issues of bias, fairness and ethicality in language testing] within the framework of reliability and validity", but that most of this discussion had previously been concerned with the test itself. She went on to explain that there had begun a shift in focus from the test itself to the consequences of test use. Bachman (2000: 23) pointed out, "There can be no consideration of ethics without validity. At the same time, investigating the construct validity of interpretations without also considering values and consequences is a barren exercise inside the psychometric test-tube, isolated from the real-world decisions that need to be made and the societal, political and educational mandates that impel them". Because the various consequences of different kinds of assessments form "part of the more recent approaches to [issues of] validity, ethical concerns of potential harm and fairness need to be examined with more than test-internal estimations of reliability and bias" (Lynch, 1997: 317).

Bachman (2000: 25) assures us that "[v]alidity and fairness are issues that are at the heart of how we define ourselves as professionals, not only as language testers, but also as applied linguists". This is echoed in the seventh principle of the International Language Testing Association's (ILTA) Code of Ethics (http://www.iltaonline.com/code.pdf; last accessed 6/2/11) which states, "Language testers in their societal roles shall strive to improve the quality of language testing, assessment and teaching services, promote the just allocation of those services and contribute to the education of society regarding language learning and language proficiency". How do we reconcile the utilitarian, instrumentary motivations of students to learn English, with our professional motivation to maintain high ethical standards? The intention of testing professionals as described by Bachman is a reflection of humanistic ethics: a sincere desire to do what's best for society. Without delving deeply into theories of the perception of humanistic ethics, we can say that the outlook of this institution with regards to the ethics of assessment is what Alan Davies (1997: 235) described as 'teleological – "do it because it will get the best results", which is based on a utilitarian form of ethics propounded by the philosopher John Stuart Mill (as opposed to the deontological ethics of Immanuel Kant – "do it because it is inherently right"). This statement is supported by the fact that there is great pressure for the college to generate palpable and measurable programme effectiveness in market terms. This was a concern of Davies (1997: 236) when he said, "[T]he increase in commercial and market forces, as well as the widespread use of language assessment as an instrument in government policy, may pressure language testers into dangerous and unethical conduct". The bottom line is that students depend on the instructors and administration to provide them with the knowledge and skills they need to enable them to do well in the external examinations they take, and the internal marks they receive, in order to gain either admission to the Bachelors programme or lucrative employment, or indeed both! For them, this is a good reason for the dependence upon formal and summative types of assessment. Viewing assessment through the lens of teleological ethics implies that the ends justify the means.

In an overview of assessment practices, Black (2001: 65) stated, ""As reformers dream about changing education for the better they almost always see a need to include assessment and testing in their plans and frequently see them as the main instruments of their reforms. This is because assessment and testing are both ways of expressing [outcomes] and means to promote or impose them". As Stiggins (2002: 1) pointed out, "When it comes to assessment, we have been trying to find answers to the wrong questions. Politicians routinely ask, 'How can we use assessment as the basis for doling out rewards and punishments to increase teacher and student effort?' They want to know how we can intensify the intimidation associated with annual testing so as to force greater achievement". Shohamy (2000: 2) referred to this 'intimidation' as "using tests as disciplinary tools", explaining that "since it is realised that test takers will change their behaviour in order to succeed on tests, those in authority will use them as a means to cause changes in behavior in accordance with their priorities and criteria". However, as experts have noted (Shohamy, 2000; Black, 2001; Bachman, 2000), such 'achievement' is an illusory one; and to exacerbate matters, this is not well-understood by laypeople. Shohamy (2000: 8) noted that the "main characteristics [of tested knowledge] are that it is narrow, simplistic and often in contradiction to experts' knowledge. After all, the information included on tests is only a *representation* of real knowledge; it is monological, based on one instrument (a test), on one occasion, detached from meaningful context and usually with no feedback given to test takers for improvement" (italics mine). This sentiment was reiterated by Black (2001: 70): "The standards dream was a straightforward vision of a way to meet both these needs: the targets would be set and promulgated and external tests would show whether or not they were being achieved. … The attraction of this approach is its simple appeal, particularly to those not in directly touch with the complexities of schooling".

Ethics of the Gate-keeping Functions of Tests

"Perhaps the first language tester to voice concerns in print about the ethics of test use and the political purposes language tests may be used to serve, [explicitly linking validity itself to test use], was Spolsky (1981, in Bachman, 2000: 16), who pointed out the language tester's post-modern predicament: each use of a language test for a gate-keeping or selection decision offers a choice of options, all of which are imperfect and fraught with problems". Later, Spolsky (1997: 242) reviewed what has been learnt in the past hundred years with regards to the ethics of gate-keeping tests. He focused on "the use of examination results to determine qualifications for positions or for training for positions" (Spolsky, 1997: 242). He pointed out, perhaps obviously, that the use of examinations in gate-keeping functions like university admission carries with it a high individual cost, and that we need "to be troubled by how much confidence to place in using results like these in making fateful decisions about people (Spolsky, 1997: 246). He stressed that even though test results may be useful in making admissions decisions, they "must not be left as the sole arbiter", and that admissions processes must "remain under responsible and challengeable human control" (Spolsky, 1997: 247).

Ethics of Teaching to the Test

'Teaching to the test' is a direct result of such intimidation. Hamp-Lyons (1997: 296) reported that in the USA, "[T]eachers said they felt pressure to teach test-taking skills, to focus on topics known to be on the test and to begin test preparation more than a month before the test…" The same sort of behaviour was reported in the Gulf News, an English language daily newspaper in the UAE (http://gulfnews.com/news/gulf/uae/education/english-proficiency-test-is-likely-for-grades-10-and-11-1.167919, 3/24/2007, last accessed 12/9/11), quoting a grade 12 student as saying that "they were required to learn up to 2000 words and place them in coherent sentences". More evidence of the intimidating or disciplinary effect national exams have may be gleaned from another article in different UAE newspaper, The National, which revealed that a report published in May, 2009 (http://www.thenational.ae/news/uae-news/education/large-numbers-still-need-extra-english-tuition-before-university, last accessed 6/11/10) by the Knowledge and Human Development Authority (of the UAE) "suggested that teaching methodology might be behind low English-language attainment in schools using the MoE curriculum, [and that] English-language teachers in state schools were sometimes not the 'best role models' because they did not speak the language fluently".

There is the question of what CEPA scores are actually being used for. "Shohamy (1993a; 1993b; 1997) has argued that language tests generally serve ethically questionable and unstated political purposes that are often quite distinct from their stated purposes" (Bachman, 2000: 16). Additionally, the newspaper article quoted above clearly shows that in the UAE at least, the onus of responsibility for students' academic success is seen as primarily upon their teachers. In such situations, the test becomes the de facto curriculum, which consequently narrows it (as noted by Hamp-Lyons, 1997, and many others). With this narrowing of the curriculum, and pressure to prepare students for the CEPA, it is not surprising that there has been "a mass increase in achievement", as Dr Annie Brown (former director of CEPA) was quoted as saying in 2009 (http://www.thenational.ae/news/uae-news/education/large-numbers-still-need-extra-english-tuition-before-university, last accessed 6/11/10). But Black (2001: 72) cautioned us against becoming too optimistic about such claims, pointing out that "whilst [there may be improvement] in pupils' test performance, standards imposed by the pressure of external tests can be counterproductive in that they can damage classroom teaching and learning, [and therefore such gains may be] short-term and illusory". Of even greater concern for the CEPA test and other tests used as admissions criteria for which students cram and are coached, is the observation made by Christie (1995: 112): "In these circumstances, teaching to the test is the only sensible strategy but, as Goodhart's law demonstrates in economics, attempts to manipulate a predictive variable, rather than to manipulate the performance that the variable predicts, immediately destroy the set of relationships which lent the variable predictive validity in the first place."

So far, a brief overview of testing ethics and reflections upon the practical realities facing the profession's stated aim of testing fairness presents a gloomy picture. However there are voices of optimism. Lynch (1997: 318) stated that "it is not the act of testing that is to blame. Tests are seen as efficient means for identifying differences in ability that exist. They do not create those differences." Bachman (2000: 23) posited that "our future as a profession lies in avoiding the Charybdis of obsessively chasing constructs, on the one hand, and the Scylla of born-again faith in ethics, on the other. ...[T]his can be best achieved by working within a conceptual framework that is broad enough and sufficiently flexible to encompass the traditional measurement qualities of reliability and validity, on the one hand, and considerations of impact – the consequences and ethics of test use – on the other." Howe (1994: 27) stated that "so long as individuals are afforded equal opportunities to obtain an education, inequalities in educational results are morally permissible." This is nevertheless the ethics of teleology. And as practical as these voices seem to be, it is difficult to imagine how we might ever be able to truly account for all 'considerations of impact' if we view the role of testing as "linked to social structures", as Hamp-Lyons has suggested. She asks us consider that "the impact of assessment and the expectations laid upon it spread all the way out to the society as a whole; it is not only test developers whose work has 'impact'. It is also testing agencies who make policy and economic decisions about the kinds of testing to support and the kinds that will not be supported; it is textbook publishers, who make economic decisions about the kinds of textbooks teachers and parents will buy to 'ensure' their children are ready for the test; it is school districts, boards and ministries of education, and national or federal governments who bow to pressure to account for the progress of pupils and the value-added effect of education" (Hamp-Lyons, 1997: 298). Alderson (1991: 66) reflected twenty years ago that, "tests and examinations are frequently significant barriers to educational opportunity, to economic and social advancement, to travel, and so on, as well as to fun".

2.13 The Affective Domain of Test Takers

In the year 2000, Beeston (in Fulcher and Davidson, 2007: 127) quoted the University of Cambridge ESL tests as saying, "For UCLES EFL, [quality assurance] starts with ensuring that we know all about the different kinds of people who take our examinations and exactly what it is they need and expect when they enter an examination." However, Bachman (2004: 50) asserted that "[b]ecause of the complexity of and interactions among various components of language ability and factors in the testing procedure, it is not possible for us to specify all the factors that affect test performance". Broadfoot (2005: 135) went further, referring "to the inseparability of the affective and cognitive domains in learning and, hence, of the corresponding need to take both into account in the design, conduct and interpretation of assessment.. Moreover, in terms of Messick's other dimensions of relevance, utility and social impact, the need to consider affect becomes overwhelming". Bachman (2008: 38) also asserted that "[p]erhaps the greatest source of [test] subjectivity is the test taker herself who must make an uncountable number of subjective decisions, both consciously and subconsciously, in the process of taking a test".

Therefore, even though it is true that we cannot possibly account for all the factors that affect test performance, attempts have been made to measure factors considered important and/or influential. SLA research "spurred language testers to investigate field independence/ dependence (Stansfield and Hansen, 1983; Hansen, 1984; Chapelle, 1988), academic discipline and background knowledge (Erickson and Molly, 1983; Alderson and Urquhart, 1985; Hale, 1988) [and] the strategies involved in the process of test-taking itself (Cohen, 1987)" (Bachman, 2000: 3). On the characteristics of test takers: "Kunnan (1998a) lists over 20 studies investigating test-taker characteristics, such as academic background, native language, culture, gender and field dependence. Other characteristics that have been studied include occupation (Hill, 1993), aptitude (Sasaki, 1996; Sparks et al., 1998), background knowledge (Clapham, 1993, 1996), and personality characteristics (Berry, 1993)" (Bachman, 2000: 11). Additionally, Broadfoot (2005: 131) pointed out that "commonsense alone is sufficient to tell us that if a learner likes and/or respects a teacher, if they are in a supportive group of peers, if the culture in the classroom is conducive to learning and, above all, how they see their own strengths and weaknesses, are factors that are likely to play a key role in the engagement and motivation of the individual concerned". Bachman organised this sort of research into 3 broad categories:

> "1) Characteristics of the testing procedure, including raters;
> 2) The processes and strategies used by test takers in responding to test tasks;
> 3) The characteristics of the test takers themselves." (Bachman, 2000: 10)

He explained that "[a] major concern in the design and development of language tests is to minimize the effects of test method, personal attributes that are not part of language ability, and random factors on test performance" (Bachman, 2008: 166). The difficulty in doing this lies in our very humanity, which we rely upon to develop assessments. "In reality, every stage of the designing, constructing, managing and administering of assessment instruments – as well as the use of the data to make judgements about the possession of certain knowledge, skills and dispositions – involves human intervention and hence values that, separately and collectively, influence every stage of the assessment process" (Broadfoot, 2005: 129). As Aldous Huxley (in Fulcher and Davidson, 2007: 127) said, "Consistency is contrary to nature, contrary to life. The only completely consistent people are the dead."

2.14 Comparable Studies

2.14.1 Recent Predictive Validity Studies

Predictive validity studies in the past 10-15 years (as well as earlier) have primarily focused on the correlative comparison between students' results on one or more English proficiency tests (i.e. TOEFL, IELTS) and these students' subsequent GPAs (Bannerjee, 2003; Feast, 2002; Lee & Greene, 2007; Kerstjens & Nery, 2000; Hill, Storch & Lynch, 1999; Cope, 2011). Most predictive validity studies report weak (r= <0.30) or results considered 'sufficient' for validation and/or admissions purposes (r= >0.30-0.50) (Davies, 1988; Bellingham, 1993; Feast, 2002; Cope, 2011). "Examination agencies such as the Educational Testing Service (who are responsible for the TOEFL) argue that the definition of 'sufficient language proficiency' varies from institution and is largely determined by each institution's needs and support facilities. In order to take these specific institutional circumstances into account, they recommend that institutions conduct their own predictive validity studies" (quoted in Banerjee, 2003: 8).

Many studies have relied upon a combination of quantitative data that includes inferential statistics (correlations, regression analyses) and descriptive statistics (grades, academic averages) (Lee & Greene, 2007; Cope, 2011). With the increasing acceptability of mixed method (MM) research designs, predictive validity studies have been enhanced by additional and informational qualitative data, usually in the form of questionnaires and interviews (Lee & Greene, 2007; Cope, 2011). As noted by Lee and Greene (2007: 387), "...although it is generally accepted that language proficiency alone, as measured by a predictor such as [a test] score, is no guarantee of academic success, this mixed method study generated specific instances of other important factors and their possible mediating roles in the predictor-achievement relationship".

The fact that academic performance is necessarily influenced by personal, individual, non-linguistic factors is a limitation that almost all of these researchers acknowledge. As Broadfoot (2005: 139) noted, "...what we can and should do is to recognize that all learners are first and foremost sentient beings and, hence, that the quality and scope of their learning is likely to be at least as closely related to their feelings and beliefs about it as it is to their intellectual capacity". (The collection and collation of questionnaire and interview data in the present study was an attempt to address this noted shortcoming.)

The findings of several predictive studies are summarised below, and in Table 2.3 that follows (along with summarized data from the current study, for the purpose of comparison).

Bellingham, L. (1993). *The relationship of language proficiency to academic success of international students.* New Zealand Journal of Educational Studies, 30 (2) (p. 229-232)
N= 38
Data: IELTS and GPAs
Positive correlation: r=0.52

The author "[i]nvestigated the relationship between international students' English-language proficiency and first-semester GPA in the national certificate of business program at New Zealand Polytechnic. … [He] attributed the more significant correlation coefficients found in his study, compared with those in previous studies, to two factors: a relatively homogeneous academic environment and a wide range of IELTS scores." (Lee & Greene, 2007: 367)

Cotton, F. & Conrow, F. (1995). *An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania*. IELTS Research Reports (Vol. 1). Canberra: IELTS Australia. (p. 72-115)

N=33 (international students studying in Australia)

Data: IELTS scores and GPAs, staff ratings of academic achievement and students' self-ratings of performance.

No positive overall correlations were found. Correlations for reading and writing subtest scores were as follows: staff ratings of academic achievement: r=0.36 and 0.34, and students' self-ratings of performance: r= 46 and 0.39.

"Several key intervening variables were briefly investigated, namely the amount of English language tuition received, motivation, cultural adjustment and welfare difficulties experienced by international students." (p. 72)

Kerstjens, M. & Nery, C. (2000) *Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance*. IELTS Research Reports (Vol. 3). Canberra: IELTS Australia, p. 85-108

N=113 (2 levels – vocational & academic – of international students enrolled in an Australian tertiary institution, Faculty of Business)

Data: IELTS scores and GPAs, student questionnaires and staff interviews

In the total sample (both levels together), only significant correlations uncovered were for Reading and Writing subtests and GPA (r=0.262 & 0.204 respectively). Separately, no significant correlations were found, except the Reading score for the academic group (r=0.287). "The regression analysis found a small-to-medium predictive effect of academic performance from the IELTS scores for the total sample and the [academic] group, accounting for 8.4% and 9.1% respectively." (p. 85) "Aside from language, staff also saw sociocultural and psychological factors such as learning and educational styles, social and cultural adjustments, motivation and maturity, financial and family pressures to have an influence on the academic outcomes of international students in their first semester of study." (p. 85)

Feast, V. (2002). *The impact of IELTS scores on performance at university.* International Education Journal, 3 (4) (p. 70-85)

N=101

Data: IELTS scores and GPAs (international undergraduate students in Australia)

"A regression coefficient of +0.39 for IELTS indicates that there is a positive relationship between the IELTS score and GPA. Thus higher IELTS scores are related to higher mean GPA scores." (p.78)

Banerjee, J. (2003) Interpreting and Using Proficiency Test Scores, unpublished doctoral thesis, Lancaster University
N=25, further reduced to 8 students (international graduate students at a university in the U.K.)
Data: Student questionnaires, interviews with students, staff and admissions personnel
"Chapters 7 and 8 [about student effort and level of English competency, showed] that a student's initial language proficiency is a good predictor of the nature of the students' study experience and the nature and the severity of the 'cost' experienced." (p. 394) "'Cost' [is] defined as the additional time and effort students [need] to expend in order to cope with their studies, over and above the time and effort they [believe] a native speaker in their cohort [has] to expend to achieve the same result." (p. 9)

Lee, Y. & Greene, J. (2007) *The Predictive Validity of an ESL Placement Test.* Journal of Mixed Methods Research, Vol. 1 (4), October 2007, 366-389
N=100 (graduate students in a large, public university in the U.S.)
Data: CEEPT scores (the university's ESL test) and GPAs, student self-assessment questionnaires, interviews with students and faculty members
"Although non-significant correlations were found between test scores and GPA, qualitative findings indicated that English skills are an important factor affecting students' course performance. Additional mixed methods analyses found that variations in students' views of academic success and their relevant background knowledge can help explain the overall insignificant relationship between ESL placement test scores and GPA."

Cope, N. (2011) *Evaluating Locally-developed Language Testing: A Predictive Study of 'Direct Entry' Language Programs at an Australian University.* Australian Review of Applied Linguistics, Vol. 34, No. 1 (p. 40-59)
N=138 (international students accepted to degree programmes from English readiness programme, named "Direct Entry")
Data: language assessment scores in Direct Entry programs and GPAs
"Research objectives were seen as twofold: it was envisaged that the present study's 'grade correlation' findings, if positive, could generate summative evidence of some value in the validation of NCELTR's Direct Entry programs for international students; and, if negative, would serve formative ends through providing an empirical starting point for remedial action." (p. 46)

Table 2.3: Summary of some Predictive Validity Studies

Researcher/s & Title of Study	Sample size (N)	Source/s of Data	Correlation Results
Bellingham, L. (1993). *The relationship of language proficiency to acade-mic success of international students.* New Zealand Journal of Educational Studies, 30 (2) (p. 229-232)	N=38	IELTS and GPAs	Positive correlation: r=0.52
Cotton, F. & Conrow, F. (1995). *An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania*. IELTS Research Reports (Vol. 1). Canberra: IELTS Australia. (p. 72-115)	N=33	IELTS scores and GPAs, staff ratings of acade-mic achievement and students' self-ratings of performance	No positive overall correlations. Correlations for reading & writing subtest scores: staff ratings of academic achievement: r=0.36 & 0.34; students' self-ratings of performance: r= 46 & 0.39.
Kerstjens, M. & Nery, C. (2000) *Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance*. IELTS Research Reports (Vol. 3). Canberra: IELTS Australia. 85-108	N=113	IELTS scores and GPAs, student questionnaires, staff interviews	In total sample, only significant correlations for Reading & Writing subtests and GPA (r=0.262 & 0.204 respectively). Separately, no significant correlations except the Reading score for the academic group (r=0.287).
Feast, V. (2002). *The impact of IELTS scores on performance at university.* International Education Journal, 3 (4) (p. 70-85)	N=101	IELTS scores and GPAs (internat'l undergrad sts in Australia)	A regression coefficient of +0.39
Banerjee, J. (2003) <u>Interpreting and Using Proficiency Test Scores</u>, unpublished doctoral thesis, Lancaster University	N=25, later reduced to 8	Student questionnaires, interviews w/ sts, staff & admissions personnel	"…a student's initial language proficiency is a good predictor of the nature of the students' study experience and the nature and the severity of the 'cost' experienced."
Lee, Y. & Greene, J. (2007) *The Predictive Validity of an ESL Placement Test.* Journal of Mixed Methods Research, Vol. 1 (4), October 2007, 366-389	N=100	CEEPT scores (ESL test) & GPAs, student self-assessment questionnaires, interviews w/sts & faculty	nonsignificant correlations were found between test scores and GPA
Cope, N. (2011) *Evaluating Locally-developed Language Testing: A Predictive Study of 'Direct Entry' Language Programs at an Australian University.* Australian Review of Applied Linguistics, Vol. 34, No. 1 (p. 40-59)	N=138	language assessment scores in Direct Entry programs and GPAs	Highest correlation recorded for APP data & MU GPAs: r = 0.418, significant at the 0.05 level. The correlation for the DEEP data & MU GPAs: r = 0.361, significant at the 0.05 level
Current Study *Explorations in the Predictive Validity of a Regionally-Developed English Proficiency Exam: An Embedded Case Study*	N=352, later reduced to 225	CEPA, GCS (final secondary exam), final college exam scores, student questionnaires, group interviews with students, interviews w/staff & administration	r=0.476 for DF & r=0.523 for HD: CEPA & 1st yr college final English marks in addition to… r=0.798 for DF & r=0.469 for HD: MLR formula & 1st yr college final English marks

The current research study not only considers the predictive validity of the CEPA, but also considers the feasibility and desirability (or lack thereof) of locally-produced proficiency examinations. With this in mind, the next sub-section covers an overview of published studies about locally and regionally developed proficiency examinations.

2.14.2 Local /Regional Design and Development of High-Stakes Proficiency Examinations

There does exist a fair amount of literature which compares the assessment programmes of different countries (Broadfoot, 1992, 1993, 1996; Hawthorne, 1997; Nuttall, 1986; Tabberrer and Metais, 1997), but not much that is recent (i.e. after 2000), and this is not a focus of the current study. What follows is a review of several published studies outlining initiatives in regionally-developed English proficiency exams.

Benin

In this study, Akoha (1988, in Alderson & North, 1991: xiv) described "an all-too-familiar situation with respect to the poor quality of public language examinations. Although the setting is Benin, readers will recognise many of the problems: a desire to change the examination in order to reflect changes in the curriculum and materials, yet considerable resistance to, and uncertainty about appropriate directions for, change; … serious problems of lack of specifications, lack of control and likely poor reliability; poor or non-existent training in testing, in item writing… and pressure on examiners to give high grades". Akoha (1991: 203) talked about "the inhibitive power of tradition… In spite of the impressive progress made in testing theory and test technology in developed countries, very little has changed in Benin and perhaps in other developing countries due to numerous factors related to politics, economy and culture. Few developing countries' educational systems can afford the increasing sophistication of English examining boards in the West".

South India

Brendan Carroll (1991, in Alderson & North, 1991: 23), in his article about resistance to change regarding language examinations, described his experience in South India in the early 1970s. The British Council "set up a large programme of ELT support to Universities in South India". As he and his team advocated teaching methodologies and strategies based on stimulus-response theory and testing based on "objective, discrete item types", they met strong opposition from Indian English Departments. The teaching of English there was steeped in a desire to cover what was deemed "the best in English Literature together with examinations in which students were expected to show a critical appreciation of selected classics. In fact, the majority of students tackled their examinations by way of prepared bazaar notes and it was widely held that many of these students could barely write or speak a single correct English sentence". Carroll reflected on the experience saying that he could be philosophical about the resistance that they met, seeing it more as "a clash of cultures" than anything else. In conclusion, Carroll echoes the sentiments expressed by Akoha and others about innovation and change: "The lesson to be learnt was that *the fashionable theories of academic pundits in USA and Britain were not necessarily fit for export*" (italics mine) (ibid).

Sri Lanka

In an article entitled, *Validating Tests in Difficult Circumstances*, Wall, Clapham and Alderson (1991: 210) described attempts made to validate a new national examination of English in Sri Lanka. Developed by the Curriculum Design and Development Centre (CDDC), the intended purpose of the National Certificate of English (NCE) was "to certify relatively high levels of proficiency in English for the non-school population, as it did not relate to any syllabus or textbook within the school system".

The authors explained that the development of the NCWE was part of an initiative started by the Sri Lankan government in the 1980s to improve the standard of English, focussing mainly on primary and secondary school levels. The reported educational situation in Sri Lanka is not unlike many in the developing world, suffering as it does from strained resources. Some of the problems faced in Sri Lanka are:

1) Most teachers are poorly paid, poorly trained and often have "a poor command of the language";
2) Examinations are administered in school classrooms, which are often open to street and hallway noise and have no electricity;
3) Poor supplies ("paper is hard to come by and expensive, and printing facilities are typically rudimentary")(Wall, Clapham and Alderson, in Alderson & North, 1991: 210-211)

"In this context, the NCE was highly innovatory. It consisted of two parts. Part 1 has Reading, Writing and General Proficiency components" … and Part 2 is an oral test and consists of four components. Because of the difficulties in administering speaking tests to large numbers of candidates, Part 2 is only taken by a reduced number of candidates, those who perform 'adequately' in Part 1 (i.e. achieving 40%) (Wall, Clapham and Alderson, in Alderson & North, 1991: 211). The NCE has been administered annually since its inception in 1986. "In 1986, Lancaster University's Institute for English Language Education was asked to conduct a study of the reliability and validity of the examination." (Wall, Clapham and Alderson, in Alderson & North, 1991: 212)

"Perhaps the most difficult problem to overcome was the fact that no public examination in Sri Lanka had ever undergone the type of scrutiny that was necessary for this study. … The CDDC team, which had been commissioned to write several English examinations besides the NCE, had always been overworked and was rarely able to think about, much less act on, ideas for evaluating their own work". In 1986, the authors of this study were obliged to create systems for collecting and handling data. In 1987, they also "established less elaborate ways of conducting the most important validity and reliability studies, so that the CDDC team could do its own evaluations in the future." Some of the problems of poor working conditions, technological deficiencies and ignorance of the need to analyse exam results have been addressed. At the time of writing, several difficult problems remained: "shortage of materials; problems with the post, telephones and transportation; severe staff shortages [exacerbated] by many non-testing-related duties; and lack of training and motivation in many of their associates" (Wall, Clapham and Alderson, in Alderson & North, 1991: 223)

In this study, content, concurrent and construct validity were investigated with varying degrees of success. As indicated by the authors, the challenge remains finding the most economic means of investigating construct validity, which still gives credible results (Wall, Clapham and Alderson, in Alderson & North, 1991: 224).

Iran

Mohamed Salehy (2008: 5) investigated the validity of a high-stakes English proficiency examination developed in Iran, called the University of Tehran English Proficiency Test (UTEPT). Administered annually, this exam must be passed by all Iranian students who wish to be considered for doctoral candidacy, regardless of his/her area of intended academic research. According to the researcher, almost 10,000 students a year sit for this exam and as yet, "to the knowledge of the researcher, no in-depth study has ever been conducted regarding the validity of the UTEPT" (ibid.). The intention of the study, therefore, was primarily to ascertain the validity of this exam, and hopefully reveal weaknesses which could be resolved by the test administration committee (the body responsible for its construction and development). Of his research, Salehy (2008: 4) said, ″The current study partly follows Anderson et al.'s (1991) study and largely resorts to more robust devices in validation inquiry. Strategy-based approach as integrated with item types and item performance, multitrait-multimethod approach and finally factor analysis were used to investigate the construct validity of a language proficiency measure”.

His analysis showed that the test was valid as it was scrutinized from different angles. “It was deemed important that looking at the validation from a single angle was not adequate. Validation inquiry from multiple perspectives was very useful as different lines of evidence would give a fair account of validity evidence. The five lines of evidence were factor analysis, multi-trait multi-method, inter subtest correlation, protocol analysis and item analysis” (Salehy, 2008: 180). About the results, Salehy (2008: 181) said: “A post hoc analysis of the reading comprehension items showed that certain items lacked acceptable value in terms of acceptability and difficulty level. The test administration committee may decide to discard these items in future administrations of the test”.

Mexico: The EXAVER Project

One of the most relevant to the present study on the CEPA, the EXAVER project was undertaken by staff at the University of Veracruz, with advice from experts from the UK. This project was initiated in order to develop a series of English proficiency standardised exams for the university. The test developers were university instructors, advised and supported by Mexican and British external advisors (Adriana Abad Florescano, et al., 2011: 228). One of the primary motivations for beginning the project was the problem of verifying the English level of prospective students because of the high cost of “reputable English language examinations” (Adriana Abad Florescano, et al., 2011: 229) versus the widespread poverty of the student population. This problem became exacerbated as Mexican employers more and more sought employees who could speak English well. Because of this, there was “an increasing demand for valid and reliable ways of certifying [this] in the interests of the students and the community [as well]” (Adriana Abad Florescano, et al., 2011: 229). A contractual arrangement was finalised between the university and the British Council (with advisors from the Council, Mexico and, at the beginning, Cambridge ESOL) to train a team of EFL teachers in test design, development and validation. Over the next five years, the project evolved and expanded. The EXAVER exam was split into three different levels with different aims, ranging from the verification of achievement of a certain level of English after 2 semesters of General English, to an "accreditation requirement" (Adriana Abad Florescano, et al., 2011: 230) for those completing certain BAs and MAs. But even these aims evolved: "As time went on, EXAVER 1 became a minimum requirement for public primary school English teachers in Orizaba City, and numerous university exchange programmes now include an EXAVER examination as an application requirement. [It now serves] a range of accreditation needs in the community" (Adriana Abad Florescano, et al., 2011: 230-31)

Having studied several different models, the development team decided to link their exam to the CEFR and ALTE guidelines (Adriana Abad Florescano, et al., 2011: 232). About their accomplishments, the developers stated that they "... have now developed specifications for batteries of tests at the three proposed levels, word lists to accompany each set of specifications, pilot versions and a significant number of live tests. [In addition, there is] a well-developed item bank containing piloted/trialled and updated items for the future" (Adriana Abad Florescano, et al., 2011: 232).

After a period of collaboration, an important issue emerged about the appropriateness of the exam's content because "the experience of the external advisors was almost entirely with large-scale, international examinations" (Adriana Abad Florescano, et al., 2011: 232). These non-Mexican exam developers were accustomed to doing their best to avoid developing any items which might fall into a 'taboo' area, because the population of test-takers for international exams is extremely diverse. ('Taboo' meaning topic areas which typically have been known to be problematic in different areas of the world (i.e. womens' rights, birth control, politics, etc.), and as such, are avoided by item developers for international exams.) As the developers rightly pointed out, the EXAVER project did not have to be as limiting because "the intended population is essentially homogenous, sharing a first language and culture" (Adriana Abad Florescano, et al., 2011: 233-234).

Again, the EXAVER project evolved to meet the perceived need to respecify the three examinations. The developers decided to use "Weir's frameworks as a guide (The frameworks were developed by Weir and O'Sullivan at Roehampton and later published by Weir in 2005)" (Adriana Abad Florescano, et al., 2011: 234-235). The rationale for doing this was twofold: "The first was to familiarise the project team with every aspect of each examination (The weakness of working to a set of specifications developed by an external consultant is a lack of ownership by local testers and a subsequent tendency to drift from the intended level and focus.). The second reason for undertaking this re-specification was to ensure that the EXAVER tests were built on a sound theoretical basis, so that subsequent validation could be facilitated" (Adriana Abad Florescano, et al., 2011: 235).

As a result of this training and professionalisation (mentioned by the developers as one of the most important things to have been gained from the project (p. 236), "the contribution of the [team from the Centre for Language Assessment Research (CLARe)] grew increasingly less and less. ... [Now,] all procedures are carried out and monitored locally" (Adriana Abad Florescano, et al., 2011: 235). Additionally, in 2009, "... Universidad Veracruzana became the first Mexican institution to be placed alongside reputable international organisations on the table of accredited language test developers" by the Mexican National Educational Standards Institution (Adriana Abad Florescano, et al., 2011: 237). Another winner as a result of this project are the candidates themselves. With understandable pride, they noted that "candidates are now able to certify their knowledge and skills in English for a range of purposes, including certification and employment, through exams that are much more affordable than most international exams but are still recognised at the national level" (Adriana Abad Florescano, et al., 2011: 240).

The recommendations with which they conclude their paper are critically important to bear in mind when considering undertaking similar endeavours elsewhere: That such a project must not be seen as a short term effort (nine years and counting for the EXAVER, eight years and counting for the CEPA); it should be understood that such a project will involve costly, long-term investment; "clear and well-defined quality assurance systems" are essential to maintain a high level of quality; that this is not simply "writing a test", but developing "an entire test system"; that good teams need time to develop themselves and effectively work together; and finally, to stay focused on the fact that the test instrument being developed must recognise the culture and language of the candidates for whom it is developed to serve (Adriana Abad Florescano, et al., 2011: 242). This was actually pointed out as "the true value of the project": to "establish a model for others to follow". They conclude by asserting that "tests do not have to be international or expensive to be 'good'" (ibid.).

The United Arab Emirates: The CEPA

A brief description of the chronological development of the CEPA to the high-stakes exam it has become was provided in Chapter 1 (See pages 9-10.). Before proceeding with the investigation of the CEPA, it is important to familiarise the reader with the exam's design and content, since comparisons have been and will be made to international proficiency exams.

The CEPA English exam has three sections, all of which test language through a discrete point, multiple choice item format with four option clusters. As described on the CEPA website (http://ws2.mohesr.ae/cepa/CEPA_DET_EN.aspx?str=EN , last accessed June 16, 2012), the three sections are (1) Grammar and Vocabulary, (2) Reading and (3) Writing. The website also explains that the administration of CEPA English takes two hours and that "there is no break time between the three parts of the test". Other information provided on the CEPA website states that the first section is comprised of 45 grammar items and 40 vocabulary items. The second section, the Reading section, "consists of two descriptive or narrative texts of around 400 words in length, and one non-prose text, such as a web page or a brochure, with a total of 25 multiple-choice questions across the three texts" (ibid.). (See Appendix M for sample CEPA exams.) The third section, Writing, "consists of an essay task of between 150 and 200 words. The quality of student's writing is assessed in terms of grammar, vocabulary, spelling and content" (ibid.). In order to make sure that students understand what they are required to do for different sections, the instructions are written in both English and Arabic.

As a former item writer for the CEPA, the researcher attended training sessions and was privy to the instructions given to item writers by the exam development team. The supervisors cautioned writers about the following (copied from an emailed training workshop document, "CEPA English Item Writing Guidelines", August, 2007):

CEPA Grammar, POS [Parts of Speech], Vocabulary - MCQ Item Writing Tips

1. No idiomatic expressions.
2. For POS and Vocabulary items, stem vocabulary should be from a higher frequency level. Item writers must exercise good judgment here. If a test taker must understand a low-frequency word in order to get the item correct, the item is unacceptable.

3. All options should fit grammatically, e.g. count vs non-count, transitive vs intransitive.
4. The stem should carry enough meaning so you can guess the key.
5. Stem should not include irrelevant detail.
6. Stem should avoid religious topics and other topics that are likely to be controversial.
7. The stem can be one or two sentences.
8. Arabic, or other foreign words, should *not* be used in the stem.
9. The stem should represent a common use of the target item.
10. There can be only one key. When in doubt, throw it out.
11. The stem should be as clear and concise as possible.
12. Familiar proper nouns may be used e.g., *Dubai*, *Abu Dhabi*, *UAE*, *Zayed Primary School*, *the United States*, common Arab personal names.
13. The stem presents the target word in a "natural" semantic and grammatical context. Sample sentences in learner dictionaries and online corpora can give useful guidance but should not be directly copied."

As can be seen in sample exams on the CEPA website, many exam items include names of people and places which would be familiar to Persian Gulf Arabs. One example from the CEPA follows (More examples are included in Appendix M.):

When I lived in Saudi Arabia I __________ my family in Sharjah every summer.
A) will visit
B) visited*
C) visits
D) visiting

An article was published by CEPA staff (Brown and Jaquith, 2011) in which the system for quickly and efficiently marking the CEPA writing section was described in detail. They explained the procedures developed by their office for training raters, how the scanning of the completed essays has been managed, and explained the system they developed to email the scanned essays to trained raters (including a description of the software to account for marking extremes, either in leniency or strictness).

More information about the validity and reliability of the CEPA can be found in the interview section of Chapter 4, since one of those who agreed to be interviewed for this study was a CEPA supervisor. (See also Appendix F for a full transcription of the interview.)

2.15 Summary

In this chapter, the literature about several aspects of assessment has been covered: what the word 'test' actually means, validity as not only an important, but arguably the most important quality a test may possess, and the stages the discussion of this concept proceeded through to arrive at the concept of unified validity. This review encompassed current views of different types of validity, including the most relevant to the present study: predictive validity. Commentary on the test item types and marking procedures used in the production and administration of the CEPA were also included. Additionally, more abstract, introspective, intuitive aspects of testing in general, although not a focus of this study, were briefly touched upon (i.e. testing ethics). Finally, examples of predictive validity studies were presented, along with experiences of large-scale, locally-developed English proficiency examinations in studies which were published in the past ten years or so. This was examined to provide a context within which this study was initiated, and to obviate the knowledge gap this research attempted to fill, or at least contribute to filling.

The "Gap"

As revealed by the review of fairly recent research, almost no studies involve large numbers of students from which stronger claims may be made. In addition, an extraordinary set of circumstances aligned which enabled the study of a cohort of students for whom the college was able to offer unfettered admission to all who expressed a desire to study at the college, regardless of the CEPA score obtained. This is one area which has been considered a fatal flaw in every other predictive validity: that the sample is always (and necessarily) truncated, meaning that not all students are offered admission. For this reason, it is normally impossible to know empirically how well the students who did not score high enough on the proficiency assessment might have fared at the college.

Another gap this research study attempts to address is that very few predictive validity studies examine a group of students with very similar demographics. In most studies, the student group under investigation hails from a diverse variety of economic, cultural and educational backgrounds. This makes the investigation of variables which may contribute to academic success so complex that it is pointless. To be sure, similar demographics cannot account for all possible variables, but certainly they are more reduced and focused.

Finally, one of the main queries of this research was to explore to what extent the homogeneity of the student cohort's demographics and the regionally developed CEPA exam (with its ability to distinguish at a fairly low English-ability level) contributed to the relatively high reported positive correlations.

3: METHODOLOGY

3.1 Introduction

In this chapter, the research design of this study will be explained in detail, beginning with an overview of the research questions themselves. In this chapter, the research design of this study will be explained in detail. The focus of the main study is an English proficiency exam, named the CEPA (Common Educational Proficiency Assessment), originally developed by lecturers at the three government-subsidized universities in this country, it evolved to a high-stakes assessment tool. Students are placed into different English readiness class levels based on the results of this test (at this college, they're called Foundations courses). From humble beginnings, over a period of four years the CEPA as a high-stakes assessment now has a major effect on decisions taken by individuals responsible for placing secondary graduates in tertiary institutions in the UAE. We begin with an overview of the research questions themselves:

A) Is this locally-produced English proficiency examination a useful statement of competence and a predictor of academic achievement?

Two of the sub-questions cover clarifications about the validity of the CEPA itself, and its predictive validity "in gauging the ability of students to progress successfully in English courses". How CEPA scores figure into the admissions process will also be considered.

Sub-Questions:

1) Is the CEPA study a valid estimate of students' English ability before admission?
2) How are the results of the CEPA exam used in the college's admissions and placement decisions?
3) What is the predictive validity of this English proficiency exam in gauging the ability of students to progress successfully in English courses and Maths courses in which the medium of instruction is English?

B) What variables appear to positively correlate with students' ability to succeed academically?

The three sub-questions to this question explore issues concerning personal, motivational and attitudinal information about the student cohort.

Sub-Questions:

1) Are the ages and academic levels of the research participants homogenous?
2) Is the socio-economic background of the participants homogenous?
3) What are their motivations for, and attitudes towards, learning English of the study's participant students?

The organisation of this chapter has been loosely adapted from the design of Lincoln & Guba (1985, in Rudestam, 2001, p. 91) for qualitative research. Interestingly enough, even though these authors have "argued strongly that [qualitative and quantitative scientific methods] virtually represent different world views" (Lincoln and Guba, 1985 in McDonough and McDonough, 1997, p. 222), their suggestions for research design work well for mixed method research studies, as may be seen in the headings of this chapter.

This mixed method research study addresses the predictive validity of an institutionally-produced, English proficiency exam used for screening purposes in placing students with varying levels of English competence into college entry-level subjects taught in English. The purpose of this study is to converge both quantitative (numeric) and qualitative (primarily attitudes) data. In this design, questionnaire data was used to provide demographic and ethnographic information, and transcribed interviews were used to note attitudes in general amongst students and administrators, and specifically to detail admissions and placement procedures amongst those responsible for student admissions and placement. Later in the study, numeric data in the form of test scores, course marks and scanned questionnaire responses were separately analysed. The purpose for collecting both qualitative and quantitative data was to merge these results in a multiple linear regression analysis. The reasons for doing this were threefold:

- To provide empirical evidence to support anecdotal evidence that locally or regionally developed English proficiency exams can be a valid indicator of academic potential in general when the medium of instruction (MOI) is English
- To establish a range of performance indicators from identified variables for the purpose of augmenting admissions decisions at a particular institution
- To provide evidence or at least insight as to the comparability, in terms of effectiveness, dependability and practicality, of locally or regionally produced English proficiency exams versus internationally-recognised proficiency exams in order to inform the intellectual debate surrounding the use (or indeed misuse) of English proficiency exams

3.2 Determining the focus

3.2.1 The Paradigmatic

Being a comparatively young and developing area of inquiry and expertise, applied linguistics has suffered from fluctuations and redefinitions of terminology to the extent that clarity of meaning has occasionally been sacrificed. To clarify, Guba and Lincoln (1994: 105; re-stated in Mertens, 2003: 139) stated in their seminal article about "competing paradigms" that "[q]uestions of method are secondary to questions of paradigm, which [is] the basic belief system or worldview that guides the investigator, not only in choices of method, but in ontologically and epistemologically fundamental ways". The overall guiding philosophic paradigm of this research study is a pragmatic one. Teddlie and Tashakkori (2009: 7) stated that the "philosophical This explanation encapsulates much of what comprises the pragmatic 'worldview'. Echoing Howe, Tashakkori and Teddlie (2003: 713) define pragmatism as "a deconstructive paradigm that debunks concepts such as 'truth' and 'reality' and focuses instead on 'what works' as the truth regarding the research questions under investigation. Pragmatism rejects the either/or choices associated with the paradigm wars, advocates for the use of mixed methods in research and acknowledges that the values of the researcher play a large role in interpretation of results". Creswell and Plano Clark (2011: 41) agree with this and add that "[t]he focus [in pragmatism] is on the consequences of research, on the primary importance of the question asked..., and on the use of multiple methods of data collection to inform the problems under study. Thus, it is pluralistic and oriented toward ... practice. The orientation most often associated with MM is pragmatism..." Earlier, Howe (1988: 15, italics in the original) very interestingly pointed out that "...much of pragmatic philosophy is *deconstructive* – an attempt to get philosophers to stop taking concepts such as 'truth', 'reality' and 'conceptual scheme', turning them into superconcepts such as 'Truth', 'Reality' and 'Conceptual Scheme', and generating insoluble pseudoproblems in the process".

The ontology and epistemology of paradigms differ in fundamental ways (as previously noted). "Ontology" can be described as how reality is perceived. "Epistemology concerns the relationship between the researcher and the participant/s" (Teddlie and Tashakkori, 2009: 89). To these, in an "expanded paradigm contrast table", Teddlie and Tashakkori (2009:88) added the ways in which paradigms differ with regard to axiology (the "role of values" in the research study), and how they view the possibility of causal linkages and generalisations. Their descriptions of the pragmatic views of these dimensions may best elucidate why this is the paradigmatic focus of the present study (as illustrated on the following page in Table 3.1):

Table 3.1: Philosophical dimensions of pragmatism (Adapted from Teddlie & Tashakkori, 2009: Table 5.2, p. 88)

With regards to:		...pragmatism advocates:
Epistemology		both objective and subjective points of view, depending on the stage of research cycle
Axiology		that values are important in interpreting results
Ontology		diverse viewpoints regarding social realities; best explanations are within personal value systems
The possibility of causal links		causal relations, but they are transitory and hard to identify; both internal validity and credibility important
The possibility of making generalisations		an emphasis on ideographic statements; both external validity and transferability issues are important

3.2.2 The Methodological

When scholars discuss academic inquiry, it is often within the parameters of two general research methods: quantitative and qualitative. The debate amongst scholars as to which scientific method was more suitable for the social sciences lasted for several decades, however, "[a]s noted by Brewer and Hunter (1989: 22), most major areas of research in the social and behavioural sciences now use multiple methods as a matter of course: 'Since the fifties, the social sciences have grown tremendously. And with that growth, there is now virtually no major problem-area that is studied exclusively within one method' (Tashakkori & Teddlie, 1998: 5)." Tashakkori and Teddlie (1998: 19) put forward that research studies which combine both methodologies within different phases of the research process are more correctly named "products of the pragmatist paradigm". This can be a very useful way of describing mixed-method studies such as this one as it does not confine the researcher to one prescriptive point of reference or view. The design of this research study follows this trend, as it combines quantitative and qualitative data gathering and analysis methodologies. Far from being mutually exclusive, they are more often regarded as complementary and even compatible. As noted by Tashakkori and Teddlie (1998: 12), "[The] similarities in fundamental values [between qualitative and quantitative inquiry] include the belief in the value-ladenness of inquiry, belief in the theory-ladenness of facts, belief that reality is multiple and constructed, belief in the fallibility of knowledge, and belief in the underdetermination of theory by fact."

3.2.3 The Research Method: Case Study

The term "case study" has become a ubiquitous one in the social sciences, and as a result, what is actually meant by this term has become somewhat muddled. Nunan (1992: 74) said that it "resembles ethnography in its philosophy, methods, and concern for studying phenomena in context, but is more limited in scope". Other commentators have described the case study as pre-experimental (Cohen and Manion, 1985; Yin, 1984). Topics for applying the case study method arise from at least two distinct situations. "First and most important (Shavelson & Townes, 2002: 99-106), the case study method is pertinent when [the] research addresses either a descriptive question (what happened?) or second, an explanatory question (how or why did something happen?) and aims to produce a first-hand understanding of people [in their actual circumstances]" (Yin, 2011: http://www.scribd.com/doc/37102046/Robert-Yin-Case-Study-Research). Within the field of applied linguistics, 'case study' can be "defined as an intensive, holistic description and analysis of a single entity, phenomenon or social unit. Case studies are particularistic, descriptive, and heuristic and rely heavily on inductive reasoning in handling multiple data resources" (Merriam, 1998: 26).

This research method was chosen because it is particularly suited to teacher research. Stenhouse (1988: 50) defines four different types of case studies, depending on their primary focus. It is his fourth which is most relevant to this research project: It is "classroom action research or school case studies undertaken by teachers who use their participant status as a basis on which to build skills of observation and analysis." This definition describes the type of research methodology used in this study. The advantages of the case study as a research method include the fact that it is "strong in reality, … and insights yielded by case studies can be put to immediate use for a variety of purposes including staff development, within-institution feedback, formative evaluation, and educational policy-making" (Nunan, 1992: 79).

Stenhouse (1988: 49) also pointed out that the case study method becomes more valuable when several of them can be combined around one phenomenon in order to make viable generalisations. This has been more recently described as an embedded case study. An embedded case study is a case study containing more than one sub-unit of analysis (Yin, 2003: 7). Because one of the aims of the study was to emphasize the critical importance of context in the determination of predictive validity, qualitative data-gathering tools were employed to construct a complete picture as possible of the different participants in the testing situation at the college. These different participants were essentially illuminative studies embedded in the overall case study of predictive validity. As explained by Creswell and Plano-Clark (2011; 95), there are "hybrid designs where researchers embed both quantitative and qualitative data within traditional designs or procedures. These approaches result in [design] variants, such as mixed methods case studies”. For this research project, the desire to produce a total picture of the context of the case was the rationale for collecting both quantitative and qualitative data.

3.3 Determining where and from whom data will be collected

3.3.1 The Institution Involved

The English proficiency exams used as placement tools and admissions criteria of a public educational institution was the focus of this investigation. It is a vocational college with an extensive preliminary English readiness programme for students who need to improve their English language competency in order to be granted admission to the degree course of their choice. The college’s Foundations (English readiness) programme curriculum concentrates on EAP. The pedagogical focus is on English as a foreign language. The institution was chosen because the researcher was employed there. Being a member of staff at this institution offered many advantages, not the least of which was ease of access.

3.3.2 Research Participants

The population of students whose results were investigated and compared was ethnically entirely Arab, and almost entirely in the 18-23 age group, owing to the requirements of the institution that applicants be recent UAE secondary school graduates. The college students whose placement tests, and other scores, were included in this study were all male. This was because the students involved were enrolled at the men’s campus where the researcher was employed. They were also almost exclusively Emirati nationals, since it is located in the UAE and this particular college only accepts citizens from the Gulf Cooperation Council (GCC) for admission. (The GCC includes 6 Arab nations in the Persian Gulf area – UAE, Bahrain, Saudi Arabia, Sultanate of Oman, Qatar and Kuwait – and is dedicated to fostering all sorts of cooperative efforts amongst the member nations.) The college students were a fairly homogenous group, as evidenced by their responses to the questionnaire items about economic status and educational background. (Their actual responses are examined in Chapter 4.)

Methodological issues of sampling and their related strategies were irrelevant to this study. This is because the entire cohort of first-year students was involved. The exceptions to this were arbitrary incidences, such as absence during the day of the questionnaire administration, or the exclusions the researcher was obliged to make for reasons which are discussed in the Data Presentation chapter (e.g. only students who identified themselves on the questionnaire could be included in the regression analysis of the main study). This might be considered convenience sampling since the students were indeed selected at least partially because of their convenient accessibility and proximity to the researcher. However, they do in fact represent the population under study: recently graduated, first year, Emirati, male college students with a CEPA score of 180 or lower. These are the students who are placed in English readiness programs and for whom the CEPA was originally designed, since (as will be discussed later) international English proficiency exams are not designed to discriminate well at a low level of English competency.

As for the other research participants, namely faculty members and administrators of the college who graciously participated in recorded semi-structured interviews, they were chosen for several reasons, and these reasons sometimes overlapped. They were chosen on the basis of their contact with students, their roles in admissions decisions, or their roles in the development or assigning of assessments. The Academic Coordinators, for example, are full-time faculty members who also are responsible for organizing syllabuses and overseeing the writing, or revamping of the major college exams for their levels. As a matter of fact, *all* those interviewed for the study had spent time teaching at the college, with the exception of the college counselor. The college counselor was chosen for his intimate knowledge of the college students in order to obtain yet another informed point of view about the difficulties these students face. The Registrar is obviously in charge of admissions, and is also aware of the backgrounds of the college students. The Assessment Coordinator of the Central Administration of the college possessed knowledge of the overall assessment scheme of all the 14 campuses, as well as being the one responsible for preparing reports about the overall term and exam results of all the campuses and ranking them according to different criteria (exam scores, coursework, etc). The CEPA Supervisor was chosen for his knowledge of the development and validation process of the CEPA itself.

3.3.3 Research Ethics

In discussing the importance of ethical considerations in research, McDonough and McDonough (1997: 54) said, "Ethics work... (a) to protect the validity of the research – for example, the achievement of good data by recognizing that data provided by informants is owned by them and its use is with their permission only; and (b) to protect the participants of the research by rules of confidentiality and consent to particular uses of the data". As much as possible, the researcher ensured the anonymity of the participants in the study. The confidentiality of students' identities was completely ensured by preserving student records in a secure location, and destroying the records once the different data analyses were completed.

For the student questionnaire, the following procedures were followed. Along with a written, explanatory statement (presented in English and Arabic), students' virtual anonymity was explained further orally, as was their prerogative to participate or not, immediately prior to the administration of the questionnaire. Written permission to record and transcribe interviews was obtained from all the teachers and administrators who were included in this study. The administration personnel who were interviewed were all well aware that their absolute anonymity could not be guaranteed due to the fact that only a few individuals hold such positions within the college, and those who hold them are well known. Written permission to be audio recorded was also obtained from all the students who agreed to participate in the group interviews.

Written permission to use the test scores of these students was acquired from the administrative officials responsible for the approval of research undertaken at this institution: the Director of Abu Dhabi Men's Campus of the Higher Colleges of Technology. (See Appendix H for a scanned copy of the college permission.)

3.4 Determining the phases of the research

It took approximately two years to collect the data required to investigate the focuses of this research. During the data collection phases, the methodology was revised and finalised. The planned phases of the study proceeded generally as planned, but the finalization of the study underwent several postponements owing to circumstances beyond the control of the researcher. There were basically six phases in this research study, and several stages within each phase. (These are graphically presented in Figure 3.1 on page 106.) Each phase is briefly discussed here with regards to the rationale for its inclusion. A detailed discussion of each phase and its stages is included in the Data Presentation chapter.

3.4.1 Phase One: The conceptualization of the Main Study

This phase was necessary for the articulation and planning of the research. The conceptualisation process was greatly facilitated through the use of a research grid (reproduced on the following page), which was provided by the supervisor of the research study, Scott Windeatt. The research grid was a chart with four major headings: Thesis question/s, Research Questions, Subjects (i.e. those subjects used to investigate a particular question), and Research Techniques. Each heading included questions to enable the researcher to formulate clear, coherent and cohesive explanations and descriptions of the intended research. For example, the "Thesis Question" heading required the researcher to consider the following questions: "What do you want to learn as a result of carrying out this study? Why? What is the main, overall 'thesis' question you want to answer?" The completed grid was discussed, revised and updated several times as the study progressed, and proved itself to be an excellent tool.

This phase also facilitated the process of obtaining the various permissions required to carry out the study because those involved understood and were clear about what was the focus of the study.

Table 3.2: Research Methods

Dissertation/Software Portfolio: Grid for research questions, hypotheses & research techniques

Instructions:

1. Write your research questions on separate lines.
2. Give brief details of the subjects you are going to use to investigate a particular question.

Finally write your research techniques on different lines to show which research questions they are intended to help you answer (and think carefully about why you have chosen a particular technique or techniques to answer a particular question).

Thesis question	***Research Questions***	***Subjects***	***Research techniques***	***Notes***
What do you want to learn as a result of carrying out this study? Why? *What is the main, overall 'thesis' question you want to answer?*	*What questions do you need to find the answers to in order to answer your 'thesis' question?*	*What subjects do you propose to use? Do they match your target audience? How many? Why have you chosen that number?*	*What research techniques can you use to answer your research questions?* *(Show which techniques will help you answer which questions)*	

3.4.2 Phase Two: Qualitative Data Collection

The decision was taken very early on in the conceptualisation stage of the research plan to administer a questionnaire to the entire cohort of first year Foundations students. This decision was taken for several reasons. First, the questionnaire format was seen as the best tool for collecting ethnographic data about the students. Assurances of protection of their identities should they wish to include their names was important and explicit in order to encourage students to volunteer this information in the questionnaire. Second, it was deemed desirable to explore students' perceptions of the CEPA test process, from the time they initially heard of it, to their thoughts about the test and the testing procedures after sitting for the CEPA.

The student questionnaire was administered in the second month of the semester. The main reason for this was timing: It was administered during the Muslim month of fasting: Ramadan. At this time, the college shortens its class times to 25 minutes, which, because of the brevity of class period and the fasting of the students, did not encourage the kind of dedicated effort especially required of both instructors and students at the beginning of the English readiness programme. A break from the routine was welcomed by all, and the timing of the shortened class times also facilitated the questionnaire's administration to the entire cohort in the space of a day. This also points clearly to the fact that the cultural awareness of the researcher himself may facilitate (or conversely inhibit) the progress of a research study. Another reason for beginning with the questionnaire is that the responses informed and affected the topic choices made for the two different kinds of interviews that followed. While most of the responses to the questionnaire were thereafter organised into quantitative data (the items which had four response options), there remained a substantial number of responses to open-ended questions, in particular the item about the desirability, or lack thereof, of being employed whilst in college and the final request for any additional comments. To begin with, most of this sort of response was provided in Arabic, necessitating translation. Then the responses were organised around the major areas of concern voiced by the students themselves.

Insert 3.1: The Students' Questionnaire (also included in Appendix C)

<u>أداة تحصيل البيانات لمشروعة بحث</u>

ضمان السرية والخصوصية

إن المعلومات التي تقدمونها في هذا الاستبيان سيتم استخدامها لتعزيز معلومات أخرى في
ومعرفة العوامل التي قد CEPAدراسات وأبحاث الدكتوراه حول "تقييم إجادة التحدث بالانجليزية"
. نأمل أن تكون هذه المعلومات مفيدة جداً HCTتؤثر على أداء الطلاب في كليات التقنية العليا
أيضا لكليات التقنية العليا.

إن طريقة الإجابة على هذا الاستبيان لن تؤثر البتة على علاماتكم ولا عليكم شخصيا. إذ سيتم حذف أسمائكم حالما نتأكد من صحة رقم تعريفكم لدى كليات التقنية العليا. وسيتم استخدام رقم تعريفكم فقط لمقارنة إجاباتكم مع البيانات الأخرى. وبعد استكمال البحث، سيتم محو كافة إجاباتكم.

يرجى الإجابة بأكبر قدر من الصدق والاهتمام.

التوقيع :   (السيدة / ليلى رمزي)

Data Collecting Tool for a Research Project

Assurance of Anonymity

The information you provide in this questionnaire will be used to augment other information in a doctoral research study about the CEPA and what factors there may be that affect students' performance at the Higher Colleges of Technology. It is hoped that this information will also be very useful and helpful to the HCT as well.

How you answer this questionnaire will have no effect whatsoever upon your marks nor upon you personally. Your names will be removed just as soon as we confirm that your HCT ID number is correct. And your ID number will only be used to compare your answers here to other data. Once the research is completed, all your responses will be discarded.

Please answer as honestly and as thoughtfully as you can.

Signed:

(Mrs. Laila Rumsey)

Questionnaire

Part One: Personal Data

1. How old are you?
   a. less than 18 years old
   b. 18-21 years old
   c. 22-25 years old
   d. more than 25 years old

2. My father ________.
   a. is illiterate
   b. can read and write
   c. is a high school graduate
   d. is a university graduate or more

Optional - The specific level of education he reached: ___________________________

3. My mother ________.
   a. is illiterate
   b. can read and write
   c. is a secondary school graduate
   d. is a university graduate or more

Optional - The specific level of education she reached: __________________________

الجزء الأول: البيانات الشخصية

1- كم عمرك ؟
أ. أقل من 18 عاما
ب. 18- 21 عاما
ج. 22-25 عاما
د. أكثر من 25 عاما

2- والدي __________
أ. أميٌّ
ب. يمكنه القراءة والكتابة
ج. خريج ثانوية
د. خريج جامعة أو أكثر

إختياري: وضّح المستوى التعليمي الذي وصلها والدك :
____

3- والدتي __________
أ. أميّة
ب. يمكنها القراءة والكتابة
ج. خريجة ثانوية
د. خريجة جامعة أو أكثر

إختياري: وضّح المستوى التعليمي التي وصلها والدتك :
____

4. My father is currently _________.
   a. employed
   b. unemployed
   c. retired
   d. deceased

5. My mother is currently _________.
   a. employed
   b. unemployed
   c. retired
   d. deceased

6. Where is your mother from?
   a. the UAE
   b. a Gulf country
   c. an Arab country other than the Gulf
   d. She is not an Arab.

>> Use the letters of these boxes to answer the next **three** questions:

4- حاليا والدي هو ___________
أ. يعمل
ب. لا يعمل
ج. متقاعد
د. متوفي

5- حاليا والدتي هي ___________
أ. تعمل
ب. لا تعمل
ج. متقاعدة
د. متوفية

6- ما هو أصل والدتك؟
أ. الإمارات العربية المتحدة
ب. بلد خليجي
ج. بلد عربي غير خليجي
د. ليست عربية

>> استخدم حروف الخانات التالية للإجابة على **الأسئلة الثلاثة** التالية:

A	B	C	D
driver	architect	civil servant	farmer
carpenter	accountant	pilot	fisherman
plumber	doctor	merchant	soldier / security guard
electrician	lawyer	teacher	telephone operator
car mechanic	director	policeman	computer technician
hairdresser	engineer	politician	medical technician
housewife			nurse

أ	ب	ج	د
سائق	مهندس معماري	موظف عمومي	مزارع
نجار	محاسب	طيار	صياد
سباك	طبيب	تاجر	جندي/ موظف أمن
كهربائي	محامي	معلم	مشغل هاتف
ميكانيكي سيارات	مدير	شرطي	فني كمبيوتر
كوافير/ حلاق	مهندس	سياسي	فني طبي
ربة بيت			ممرضة

7. Which group contains jobs that are most similar to the primary one that your father has (or had)?
   a. c.
   b. d.

8. Which group contains jobs that are most similar to the primary one that your mother has (or had)?
   a. c.
   b. d.

7- ما هي المجموعة التي تحتوي الوظـائف الأكثر شبها بالوظيفة الأساسية لوالـدك (حاليـا أو سابقا)؟
أ ب
ج د

8- ما هي المجموعة التي تحتوي الوظـائف الأكثر شبها بالوظيفة الأساسية لوالدتك (حاليا أو سابقا)؟
أ ب
ج د

9. Which group contains jobs that are most similar to the primary one that you yourself are aiming for?
a. c.
b. d.

10. Does your father have another job?
a. yes
b. no

If yes, what is the other job?______________

11. Does your mother have another job?
a. yes
b. no

If yes, what is the other job? ______________

12. Do you have brothers and sisters?
a. Yes, more than 10.
b. Yes, more than 5.
c. Yes, less than 5.
d. No, I don’t have.

13. Where do you stand with regards to your brothers and sisters?
a. I am the oldest.
b. I am the youngest.
c. I am not the oldest or the youngest.
d. I am an only child.

14. What is your marital status?
a. single
b. married, no children
c. married, with children
d. divorced or widowed

15. How long does it usually take you to get to college every day?
a. less than 15 minutes
b. 15 – 30 minutes
c. 30 – 45 minutes
d. more than 45 minutes

16. Where will you live while at college?
a. at home with my family
b. in a flat by myself nearer the college
c. in a flat with friends nearer the college
d. with relatives nearer the college

Part Two: Academic Information

17. What kind of high school did you attend?
a. public high school (government)
b. model high school (government)
c. private high school
d. home-schooled

9- ما هي المجموعة التي تحتوي الوظائف الأكثر شبها بالوظيفة الأساسية التي ترغب بها أنت؟
أ ب
ج د

10- هل عند والدك عمل آخر؟ (حاليا أو سابقا)
أ. نعم
ب. لا

إذا اجبت ب "نعم" ما هو هذا العمل؟ __________

11- هل عند والدتك عمل آخر؟ (حاليا أو سابقا)
أ. نعم
ب. لا

إذا اجبت ب "نعم" ما هو هذا العمل؟ __________

12- هل لديك إخوة وأخوات؟
أ. نعم، أكثر من 10
ب. نعم، أكثر من 5
ج. نعم، أقل من 5
د. لا، ليس لديّ.

13- ما هو ترتيبك بين إخوتك وأخواتك؟
أ. أنا أكبرهم
ب. أنا أصغرهم
ج. لست أكبرهم ولا أصغرهم
د. أنا الابن الوحيد

14- ما هي حالتك العائلية؟
أ. أعزب
ب. متزوج بدون أطفال
ج. متزوج ولديّ أطفال
د. مطلق أو أرمل

15- كم تستغرق من الوقت للوصول إلى الكلية؟
أ. أقل من 15 دقيقة
ب. 15 – 30 دقيقة
ج. 30- 45 دقيقة
د. أكثر من 45 دقيقة.

16- أين تقيم خلال دراستك بالكلية؟
أ. في المنزل مع أسرتي
ب. وحدي في شقة قريبة من الكلية
ج. مع أصدقاء في شقة قريبة من الكلية
د. مع أقارب يسكنون بالقرب من الكلية

الجزء الثاني: المعلومات الدراسية

17- ما هو نوع المدرسة الثانوية التي درست بها؟
أ. مدرسة ثانوية عمومية (حكومية)
ب. مدرسة ثانوية نموذجية (حكومية)
ج. مدرسة خاصة
د. دراسة منزلية

18. What was your overall final average from your high school exams?

a. 90 – 100 %
b. 75-89.9 %
c. 50-74.9 %
d. less than 50%

19. What stream did you follow in high school?

a. Science stream
b. Arts stream
c. Started with Science and then switched to Arts

20. Did your parents pressure you to study?

a. Yes, always.
b. Yes, but not so much that I felt overly pressured.
c. Almost never.
d. Only for certain subjects.
(Which subjects? ________________________________)

21. What was your CEPA English mark?

a. more than 170
b. between 140 and 169.9
c. less than 140
d. I don't know. / I didn't sit for the CEPA.

22. Did you take special, extra classes to prepare for the CEPA exam?

a. Yes, in my school.
b. Yes, outside my school.
c. No: I did not think it was important or that I needed to.
d. No: I wanted to but there weren't any extra classes available for me to take.

23. How did you feel when you sat for your high school English exam?

a. calm, relaxed and confident
b. a little nervous or worried
c. very nervous and worried
d. carefree

24. Did you feel any differently when you sat for the CEPA English exam?

a. Yes
b. No
c. About the same

If you answered yes, please explain why you felt differently for the CEPA:

18- ما هو مجموع مُعدّلك النهائي في امتحانات الثانوية؟

أ. 90- 100%
ب. 75 - 89.9%
ج. 50 - 74.9%
د. أقل من 50%

19- ما هو المقرر الذي درسته في الثانوية؟

أ. المقرر العلمي
ب. المقرر الأدبي
ج. بدأت بالمقرر العلمي ثم انتقلت إلى الأدبي

20- هل يضغط عليك والداك للدراسة؟

أ. نعم، دائما
ب. نعم، ولكن ليس بالقدر الكبير الذي يجعلني أشعر بضغط شديد
ج. تقريبا أبدا
د. فقط في مواد معينة
(ما هي ؟
....)

21- ما هي علامتك في الانجليزية ضمن تقييم إجادة التحدث بالانجليزية "CEPA" ؟

أ. أكثر من 170
ب. بين 140 و 169.9
ج. أقل من 140
د. لا أعرف أو لم أمتحن ال" CEPA"

22- هل تابعت دروسا خصوصية أو حصصا إضافية للتحضير لامتحان CEPA؟

أ. نعم، في مدرستي
ب. نعم، خارج مدرستي
ج. لا، لم أكن أعتقد بأنها هامة أو بأنني بحاجة إليها
د. لا، كنت أود ذلك لكن لم تكن هناك حصص إضافية متوفرة لي

23- كيف كان شعورك وأنت تجلس لامتحان الانجليزية في المدرسة الثانوية؟

أ. هادئ وواثق
ب. قلق ومضطرب نوعا ما
ج. قلق ومضطرب جدا
د. غير مهتم

24- هل انتابك شعور مختلف نوعاً ما عند جلوسك لامتحان CEPA؟

أ. نعم
ب. لا
ج. تقريبا نفس الشعور

اذا أجبت بـ "نعم"، يرجى شرح لماذا كان شعورك مختلفا في امتحان CEPA:

25. Do you feel that you were placed correctly at HCT?
   a. Yes, definitely.
   b. I'm not sure.
   c. No. I want to change my placement.
   d. No, but it doesn't matter that much to me

27. What college major do you hope to join next year?
   a. Engineering
   b. IT
   c. Business
   d. I don't know yet.

28. Were you given instruction about how to fill in computer answer sheets for the CEPA exam by your high school English teacher?

   a. Yes
   b. No
   c. Not sure/ Don't remember

29. Did your high school English teacher coach you about how to answer the kinds of questions you would find in the CEPA English exam?
   a. Yes. He/She spent more than 50% of our English lesson time doing this.
   b. Yes. He/She spent less than 50% of our English lesson time doing this.
   c. Yes. He/She gave us classes after school for this.
   d. No, not at all.

If you answered YES to #29 (a or b), tick all the ways in which you were coached:

___ He/She taught us better study skills.
___ He/She taught us answering strategies.
___ He/She gave us sample questions.
___ He/She gave us a mock exam/s.

30. Do you plan to work while you study this year?
Yes: __________ No: ________
Do you anticipate any difficulties doing this?
Yes: __________ No: ________
Please explain:

-25 هل تشعر بأن وضعك صحيح في كليات التقنية العليا؟
أ. نعم حتما
ب. لست متأكدا
ج. لا. أريد تغيير وضعي
د. لا. لكن لا أهتم كثيرا.

-27 ما هو المقرر الدراسي الذي تأمل دراسته في السنة القادمة؟
أ. الهندسة
ب. تقنية المعلومات
ج. الأعمال
د. لا أعرف بعد

-28 هل تم تعليمك بواسطة معلم الانجليزية في المدرسة الثانوية كيف تملأ صفحات الإجابة على امتحان CEPA على الكمبيوتر؟
أ. نعم
ب. لا
ج. لست متأكد أو لست متذكر

-29هل قام معلّم اللغة الانجليزية في المدرسة الثانوية بتعليمكم كيف تجيبون على نوع الأسئلة التي قد تجدونها في امتحان CEPA الخاص باللغة الانجليزية؟
أ. نعم. لقد أمضى أكثر من 50% من حصص الانجليزية لتعليمنا ذلك.
ب. نعم. لقد أمضى أقل من 50% من حصص الانجليزية لتعليمنا ذلك.
ج. نعم. لقد اعطانا حصص بعد الدوام المدرسي الرسمي
د. لا، إطلاقا

إذا أجبت بـ "نعم" (أ أو ب) لسؤال #29، أشـّر على كافة الطرق التي تم تعليمكم بها:
___ علّمنا أفضل مهارات الدراسة
___ علّمنا استراتيجيات الإجابة
___ أعطانا أمثلة عن الأسئلة
___ أعطانا امتحانا مشابها

30 – هل تنوي أن تجمع العمل و الدراسة هذا العام؟
نعم: ________ لا: ________

هل تتوقع أية صعوبات في ذلك الأمر؟
نعم: ________ لا: ________
يرجى شرح جوابك :

Part Three:

Your reasons for studying English

31. I want to study English because it will allow me to be more at ease with people who speak English.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

32. Studying English is important to me only because I'll need it for my future career.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

33. Studying English is important to me because it will allow me to meet and converse with more and varied people.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

34. Studying English is important to me because it will make me a more knowledgeable person.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

35. Studying English is important to me because it will enable me to better understand and appreciate English art and literature.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

36. Studying English is important to me because I think it will someday be useful in getting a good job.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

37. Studying English is important to me because I will be able to participate more freely in the

الجزء الثالث:

أغراضك من التعليم اللغة الانجليزية

31. أريد أن أدرس اللغة الانجليزية لأنها ستجعلني أكثر ارتياحا مع ناس يتكلمون اللغة الانجليزية.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

32. دراسة اللغة الانجليزية تهمني لأني سوف أحتاجها لمهنتي في المستقبل.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

33. دراسة اللغة الانجليزية تهمني لأنها ستمنحني فرصة لأقابل و أتحدث مع ناس أكثرعددا و تنوعا.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

34. دراسة اللغة الانجليزية تهمني لأنها ستجعلني انسان أكثر معرفة.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

35. دراسة اللغة الانجليزية تهمني لأنها ستجعلني أفهم و أقدر فن و أدب انجليزي أكثر.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

36. دراسة اللغة الانجليزية تهمني لأنني أظن أنها ستفيدني في الحاق بعمل في يوم ما.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

37. دراسة اللغة الانجليزية تهمني لأنني سأتمكن من المشاركة في نشاط المجموعات الثقافية الأخرى.

activities of other cultural groups.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

38. Studying English is important to me because other people will respect me more if I have a knowledge of a foreign language.

a. Strongly Agree
b. Agree
c. Disagree
d. Strongly Disagree

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

38. دراسة اللغة الانجليزية تهمني لأن الناس الآخرون سيحترمونني أكثر ان كنت أعلم لغة أجنبية.

أ. متفق بشدة
ب. متفق
ج. غير متفق
د. غير متفق بشدة

هذه نهاية الاستبيان. إذا كان لديك أي شيء تود قوله أو أي تعليق على هذا الاستبيان، يرجى كتابته أدناه (بالانجليزية أو العربية)

This is the end of the questionnaire. If you have anything else you'd like to say or comment upon about this questionnaire, please write it here: (in English or Arabic)

____
____
____
____

- end -

After the administration of the questionnaire, group interviews were organised with a select group of students. As it is not possible to perform group interviews with 350 students, a decision had to be made about how to select a representative group. Since it was important to have a vocal and articulate group, the researcher focused only on those students who wrote a response to the open-ended item at the end, further narrowed by relevant responses (For example, one student wrote, "It's my 17th birthday today!!"), and those who recorded their names on the questionnaire. This was arranged after the results had been tabulated. Each of these students was contacted individually and asked to join. 25 out of 47 students agreed to participate, and eventually 19 attended the actual discussion. This was planned in order to flesh out further explanation of points raised by students in the questionnaire, and to clarify noted general response trends.

Semi-structured interviews were planned and conducted with different members of the college's administration. The reason for conducting these interviews was to clarify the criteria being used for admissions and placement purposes, and to collect detailed information about the CEPA. The protocol questions were gleaned from the research sub-questions, and from points that Bannerjee (2003) focused upon in her research of predictive validity at Lancaster University.

Insert 3.2: Semi-Structured Interview Protocol (May also be found in Appendix E.)

Interview Venue: Time of Interview:
Name of Interviewee: Department of Interviewee:
Letter of 'No Objection' signed:

Part One: Admissions Decisions

1. What are admissions decisions based upon? Any other factors considered?
2. What is/are the supervisor's role/s?
3. What is the role of CEPA? Other things considered?
4. What would be considered a 'clear accept'? … a 'safe bet'? … a 'risk'?
5. Of all who apply, what percentage are accepted?

Part Two: Marking Scheme

6. What is the marking/weighting scheme and is this year's marking scheme the same as the previous year? If not, how has it differed and why was it changed?
7. How are marking scheme and weighting decisions taken?
8. How much of these decisions is decided by each individual college?

Part Three: 'Cost'

9. Is HCT's reputation a factor which is considered when making admissions decisions?
10. Is the added burden to teaching and administrative staff and academic support staff considered when making admissions decisions?
11. What do you think the added burdens are to students and staff if students are accepted at a lower level than are normally accepted?
12. Are there are emotional and physiological effects on students which you consider 'at risk'?
13. Do 'settling in difficulties' for students affect their English achievement?
14. … their performance?

Part Four: The Students Themselves

15. What do you think are the students' motivations for learning English?
16. What do you think are the students' attitudes towards learning English?
17. What factors would you expect to contribute to the likelihood that a student would succeed in his studies?
18. What factors would you expect to be a source of struggle for students?

⇨ SPECIFICALLY FOR ACADEMIC COORDINATORS (team leaders who are also student instructors themselves)

- Do you have any role in student admissions?
- What is your role in assessment decisions and organisation?
- Do the CEPA results generally match teachers' perceptions of what their students' abilities are?

⇨ SPECIFICALLY FOR THE COLLEGE REGISTRAR

- How does a student get in to the Higher Colleges?
- What percentage of all the students who apply do you actually accept?
- Do you think that the P.I. generally matches what students' academic abilities are?
- What is your role in the organisation of the CEPA exam?

⇨ SPECIFICALLY FOR THE CENTRAL ASSESSMENT COORDINATOR

- What is your role in assessment decisions and organisation?
- Do the CEPA results generally match teachers' and administrators' perceptions of what their students' abilities are?

- On your website, in the final exam reports, there are individual reports for each college for the correlation between coursework and exam marks. Is there a reason why this is not done for the CEPA and final exams?

⇨ **SPECIFICALLY FOR THE CEPA SUPERVISOR**

- How has NAPO validated the CEPA?
- How certain is NAPO that the CEPA is a reliable estimate of students' English ability?
- How does NAPO justify the exclusion of listening and speaking components?
- A teacher reported that the CEPA has been internationally accredited as a substitute for IELTS and TOEFL. Is there any truth to this?
- This same teacher provided an equivalency table comparing CEPA scores to IELTS and TOEFL. Do you know about this? How were these equivalences arrived upon?
- Did the CEPA become more difficult when non-nationals were allowed to take it?

3.4.3 Phase Three: Quantitative Data Collection

As previously mentioned, permission had been granted by the college to have the complete admissions files of the entire cohort for the 2007/2008 academic year, as well as a complete record of their marks for their first year in the two Foundations Programmes. In addition, students' responses to questionnaire items were quantified. It was possible to do this since the response choices had been restricted to an adapted Likert-like scale which included only 4 options. Doing this enabled the questionnaires to be economically scanned and marked electronically. The results were entered onto a spreadsheet. Having the questionnaire quantified greatly facilitated the analysis of the results.

3.4.4 Phase Four: Merging of Results

Pearson's correlation coefficient analysis was chosen as one of the two most important statistical analyses performed on the data collected because it demonstrates how strong the relationship is between 2 sets of variables, and whether that relationship is linear or not. It was used to explore the relationship between the CEPA results and other student scores. The verification of linear relationships between different variables was important because the final quantitative data tool, the multiple linear regression (MLR), relies on the assumption that the relationship between variables is linear. This analytical tool was chosen for its ability to predict performance indicators, and its ability to compare many variables of differing kinds. A Kolmogorov-Smirnov test was performed first to confirm the normal distribution of the dependent variables in the two MLRs performed. Responses to questionnaire items about social, educational and economic issues were included in the MLR. The MLR was chosen to statistically include and analyse the responses to the student questionnaire, their CEPA results and their final marks at the end of their foundation year. The rationale for choosing the MLR is that it allows us to compare quantitative and qualitative data and to make predictions. It also enables the comparison of a multitude of variables. Because part of the research explored factors that students themselves, as well as faculty and administration members, identified as affecting students' ability to succeed academically, the final course results were the dependent (constant) variables. The independent variables analysed were most of the quantified responses in the students' questionnaires, their CEPA scores and their high school English final marks. Variables excluded from the data collected were either those that were deemed irrelevant (according to the response pattern and the results of the student group discussions) or those which were related to the CEPA itself.

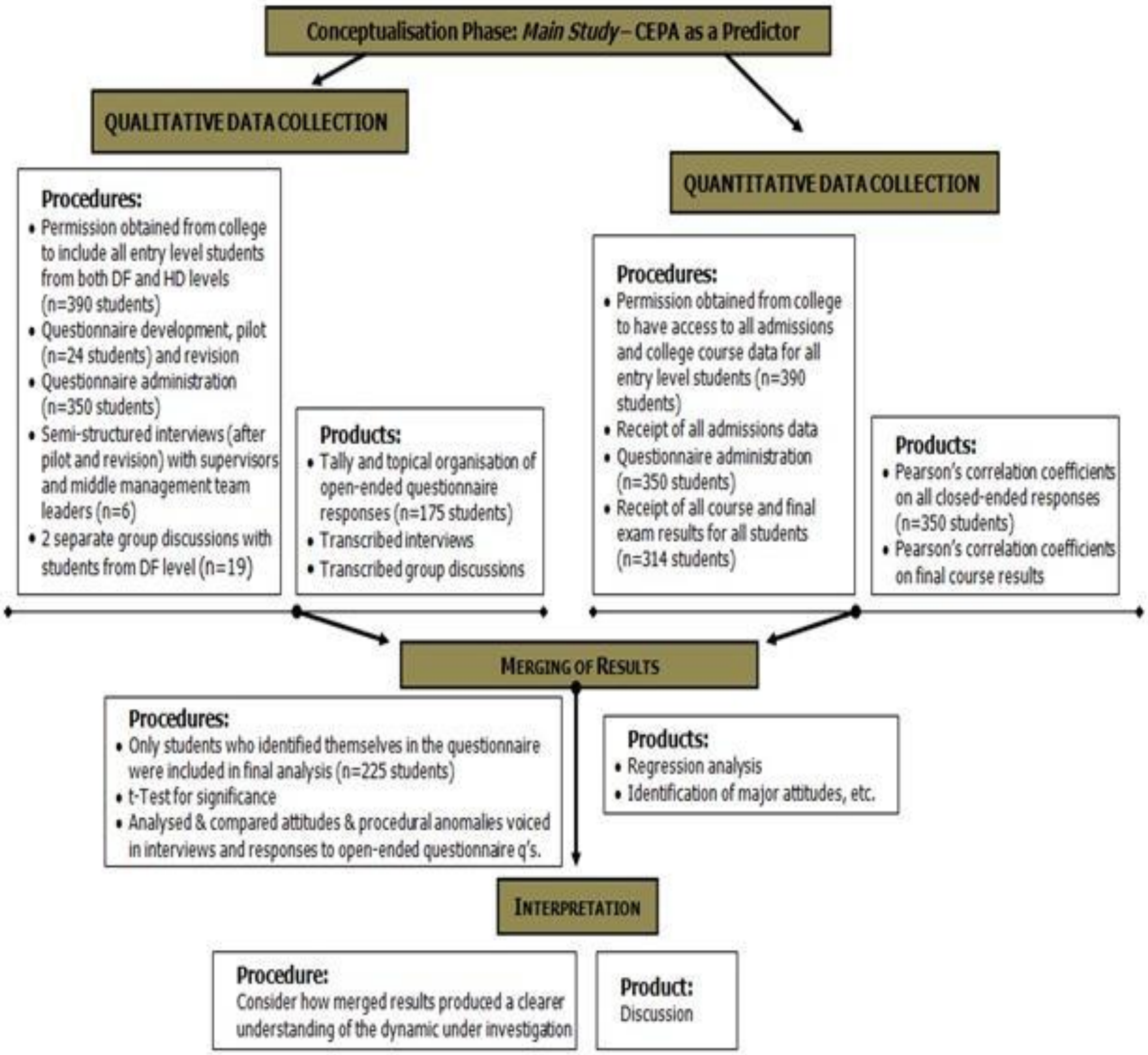
Only data from AY 07/08 was included because that was the year the student questionnaire was administered. The STEPWISE regression analysis was chosen because this analyses the effect of variables upon each other. It produces a list of factors which it has found to be statistically significant "by placing all predictors in the model and then calculating the contribution of each one by looking at the significance value of the t-test for each predictor" (Field, 2005: 161). A removal criterion automatically removes any predictor that does not prove to be statistically significant, and then the model is re-estimated for all the remaining predictors. For example, this analysis revealed that the amount of CEPA coaching a student received was significant. This list of factors can then be used to develop a model to test the reliability of the MLR results. The basic MLR model is: "$Y_1 = (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \mathcal{E}_i$. Y is the outcome variable [- the constant], b_1 is the coefficient of the first predictor (X_1), b_2 is the coefficient of the second predictor (X_2), b_n is the coefficient of the *n*th predictor (X_n), and is the difference between the predicted and the observed value of Y for the *i*th participant" (Field, 2005: 157). A Pearson's Correlation was performed on 10% of the total sample (left out of the MLR). This was a straight comparison between the students' actual CEPA scores and what the MLR model predicted their CEPA scores would be, given the variables for each student. This is the analysis that revealed the strength of the predictive validity of the CEPA for this cohort, based on the factors chosen by the MLR as most correlational.

3.4.5 Phase Five: Interpretation

In this phase, the final discussion and analysis of the research study produced some confirmatory and some non-confirmatory conclusions. This will be discussed later on in

Chapter 5. The research analysis was organised around the research sub-questions. The discussion was necessarily restricted to address the issues investigated in the study. The more general issues concerned with assessment and testing in general were discussed, as much as possible, in Chapter 2. Diagram 3.1 on the following page was developed by the researcher to graphically represent the planned organisation of the current study. The design was adapted from organisational graphics presented in Cresswell and Plano Clark (2011: 118) and Tashakkori and Teddlie (2009: 154-155).

Figure 3.1: Diagram of the Convergent Sequential Design of the Research Study


Conceptualisation Phase: *Main Study* – CEPA as a Predictor
QUALITATIVE DATA COLLECTION
Procedures:
• Permission obtained from college to include all entry level students from both DF and HD levels (n=390 students)
• Questionnaire development, pilot (n=24 students) and revision
• Questionnaire administration (n=350 students)
• Semi-structured interviews (after pilot and revision) with supervisors and middle management team leaders (n=6)
• 2 separate group discussions with students from DF level (n=19)
Products:
• Tally and topical organisation of open-ended questionnaire responses (n=175 students)
• Transcribed interviews
• Transcribed group discussions
QUANTITATIVE DATA COLLECTION
Procedures:
• Permission obtained from college to have access to all admissions and college course data for all entry level students (n=390 students)
• Receipt of all admissions data
• Questionnaire administration (n=350 students)
• Receipt of all course and final exam results for all students (n=314 students)
Products:
• Pearson's correlation coefficients on all closed-ended responses (n=350 students)
• Pearson's correlation coefficients on final course results
MERGING OF RESULTS
Procedures:
• Only students who identified themselves in the questionnaire were included in final analysis (n=225 students)
• t-Test for significance
• Analysed & compared attitudes & procedural anomalies voiced in interviews and responses to open-ended questionnaire q's.
Products:
• Regression analysis
• Identification of major attitudes, etc.
INTERPRETATION
Procedure:
Consider how merged results produced a clearer understanding of the dynamic under investigation
Product:
Discussion

3.5 Deciding instrumentation

3.5.1 Qualitative Data Collection

"Qualitative methods may be most simply and parsimoniously defined as the techniques associated with the gathering, analysis, interpretation and presentation of narrative information" (Teddlie & Tashakkori, 2009:343). This is, of course, the sort of qualitative data which was sought after and collected: questionnaires, interviews and group discussions.

Questionnaires

Questionnaires were chosen as an appropriate instrument for gathering data not only because of the large number of students (n= 347 students for the college), and ease of administration in terms of time and manageability, but also for its quantifiability. Closed-ended, four-option items formed the bulk of the questionnaire (whose 4 response options consisted of a revised Likert-type scale). This was done in order to take advantage of the offer made by the Admissions Office to scan the responses to these items and electronically organise and tally them, thereby greatly reducing the effort which would have been required to do this. Additionally, the questionnaire was divided into 3 sub-sections with separate headings to identify the focus of particular sub-sets of questions whose answers this study deals with: Part One – Personal Data (items 1-16), Part Two – Academic Information (items 17-30) and Part Three – Reasons for Studying English (items 31-38). It is also noted here (and on page 5) that the terms secondary, secondary school and high school are here used synonymously. Before organising the pilot of the questionnaire, an official translation into Arabic was prepared, as well as a soft copy of the translation arranged for, and this greatly facilitated the preparation of the bilingual presentation of the questionnaire. (See Appendix J for a scanned copy of the official translation.)

The Pilot of the College Questionnaire

The questionnaire was piloted with two sections (out of 6) of students who were at the same level as the first year cohort, but who were off-sequence students (meaning they'd begun their foundations year in January, not August). These two sections were chosen for convenience as they were students the researcher was teaching at the time. On the basis of the informal group discussions with these participant students immediately following the questionnaire pilot, their feedback was noted and some valuable alterations were made. In the pilot discussion, students made two minor corrections to the Arabic translation. With regard to the very personal questions planned for inclusion in the questionnaire, the researcher, sensitive to the strong feelings of Emirati families with regards to the protection of families' privacy, directly questioned the pilot groups about the appropriateness of these questions, fearing them too invasive. Without exception, the students felt the questions were appropriate, and some said they were not probing enough. Questions 6 (about the nationality of the mother of the student), 15 (about the commuting time to college), 25 (about their placement level at the college) and 30 (about working while at college) were added at their request as a result of our discussions, including the wording and the response options. These students also felt that there should be more choice options for questions 2 and 3 about parents' education. When it was explained that item choices had to be necessarily restricted to 4 options because bubble sheets were to be used, they specifically requested that a space be made for students to elucidate, should they so desire, and this was done. (The illiteracy rate was fairly high in the UAE until recently - 65.4% in 1980 according to UNESCO (http://unesdoc.unesco.org/images/0014/001462/146282e.pdf, last accessed 30/6/11) - and students are understandably proud of parents who have completed some level of education.)

Input from others for the Questionnaire

In addition, two Arabic-speaking faculty members read the questionnaire and made cogent and constructive suggestions to improve it, especially with regards to the Arabic translation, and the wording of two items which were potentially emotive for students. Professor Peter Tymms of Durham University also suggested in a meeting with him (August, 2007) that an item about the employment of the students' parents was important, and so an item from the Yellis © assessment of the CEM Centre was adapted for the questionnaire (items #7, 8 and 9). Finally, after the questionnaire was shared with Professor Vivian Cook, eight questions about students' motivation for learning English were inserted, using a shortened version of Gardner's (1985: appendix) motivation questions from Professor Cook's website (http://homepage.ntlworld.com/vivian.c/SLA/MotTest.htm, accessed 29 July, 2007) on his suggestion (items #31-38 in the questionnaire).

Semi-Structured Interviews

Semi-structured interviews were conducted as part of the data collection process. The interview outline was piloted with two volunteers from the faculty. Some minor changes were made and then the interview appointments were arranged. In determining if there was a measurable relationship between the predictive validity of the CEPA and the factors which may positively correlate with student success, the information obtained in these interviews proved to be very valuable. This was especially true in the interviews with the college supervisors and the college Registrar because not only were they intimately involved in the assessment of students, but they also had experience in placing students into the two different programme tracks (one of which – HD – offered better opportunities for advancement than the other). Their responses, and additionally the responses of the academic co-ordinators, helped to inform the determination of the validity, reliability and manageability of different types of assessment used at the college, as well as to what uses examinations were being put.

All the interviewees were briefed ahead of time as to the content of the interview and assured of relative anonymity prior to the interview itself. ("Relative" in the sense that absolute anonymity could not have been ethically guaranteed, as was previously mentioned.) Additionally, all interviewees agreed to be audio recorded and signed documentation to that effect. The recordings facilitated the correct transcription of these interviews.

The interview format was preferred over other formats for several reasons. First, because the interviewees were colleagues, a personalised format was deemed more appropriate. A semi-structured, quasi-formal interview style was chosen over other styles in order to enable the interviewee to elaborate, thereby providing the possibility of greater input. This was not only because questionnaires tend to constrain elaboration, but also because, as professionals and fluent English speakers, they were more likely and more able to elaborate. As Judith Bell (1993: 91) stated, "A major advantage of the interview is its adaptability". Second, in the interest of triangulation, a qualitative data collection format other than questionnaires was chosen in order to ensure that the process would be more comprehensive.

Group Interviews/ Discussions

Group interviews are often used in qualitative research as a follow-up data collection method, as it is in this research project. A group interview format was chosen to follow-up on the students' responses to the questionnaire about the test-retest research activity in the preliminary study, as well as after the administration and collation of the college questionnaire. Merton et al. (in Denzin & Lincoln, 1994: 365) called such groups 'focus groups' and the term was applied to interviews of groups after a considerable amount of research had already been completed. It was felt that an atmosphere which encouraged collegiality would be more effective at throwing light upon certain unexpected outcomes revealed in the test-retest exercise and the questionnaire. Such interviews "can yield rich material and can often put flesh on the bones of questionnaire responses" (Bell, 1993: 27). With certain issues, this proved to be the case. Ideally, the group size for such an interview should be in the range of 7-12 people, but Stewart and Shamsadani (1990: 100) have stated that the appropriate number of participants would depend on the objectives of the research. Their responses were recorded, with their permission (as previously noted), and transcribed in order to facilitate the analysis of the discussion. The goal of this analysis was to single out strongly-held majority opinions, as well as to identify patterns that recurred with both groups.

3.5.2 Quantitative Data Collection

"Quantitative methods may be most simply and parsimoniously defined as the techniques associated with the gathering, analysis, interpretation and presentation of numerical information" (Teddlie & Tashakkori, 2009: 343).

Several analytical tests were conducted in the main research study, initially this was done to compare the results within each level: Diploma Foundations (DF: the so-called 'weaker' group), and Higher Diploma Foundations (HD). Eventually the overall patterns between the two groups were compared, as well as the comparison which resulted from the merging of the analysed qualitative data and the tabulated quantitative data. The raw data was entered into analysis software for later statistical analysis. The names of participants were preserved and used only in order to enable the comparison of results in the analysis phase. Then those names were discarded.

As a result of the experience of the preliminary study (which – as mentioned – was prematurely cut off), it was clear that the numerical data collected in the main study would need to be correlated. As one of the primary aims of this study was to determine the predictive validity of the CEPA, it was necessary to determine the strength of the comparisons between the college students' CEPA results and their college course marks. The Pearson's correlation coefficient analysis tool was chosen because it "is a statistic that is calculated from data that summarizes the strength and direction of the relationship between two variables" (Bachman, 2004: 84). A correlation coefficient, also known as the Pearson product moment correlation, is an analysis "which gives a description of the overall degree of agreement between two sets of scores" (McDonough 1997: 146). Using scores plotted on a scatter-gram is a fairly straightforward statistical operation. The rationale for including scatterplot diagrams of the results is that they present a graphic visual of the correlation analysis. The correlation coefficient can have any value between –1.0 and +1.0, but a high correlation will approach +1. Pearson's correlations assume, amongst other things, that both variables are normally distributed and the x and y are independent of each other.

The other two statistical analyses were decided upon as the research progressed, as a result of discussions with a statistician from the researcher's university. While correlation analyses are useful to reveal the strength of a comparison between two variables, it remains not only limited in its scope by its two-dimensionality, it also cannot compare and analyse relationships for numerical and non-numerical data, and it should not be used to make predictions about variables (i.e. It cannot reveal any causal relationships). "The correlation coefficient is merely a measure of whether two variables are related. It does *not* indicate whether one variable causes the other" (Vernoy & Kyle,2003: 160, italics in the original). Before proceeding with the final stage of analysis, a test of normality was necessary to establish that the dependent variable is normally distributed. "The Kolmogorov-Smirnov test … can be used to test the hypothesis that the distribution is normal" (Eliott and Woodward, : 25). As clarified by the university's statistician, "If you have a larger sample size greater than 50, look at the Kolmogorov-Smirnov to make your conclusion [about normality of data distribution]. If you have a small sample size, less than 50, look at the Shapiro-Wilk to make your conclusion" (Kometa, email correspondence: 27/10/08).

The multiple linear regression was the final analysis added to the research design. The reason for its addition was its ability to compare numerical and non-numerical variables, as well as make predictions. It was also a logical choice since the vital assumption one must make in using this type of analysis is "that the variables used to make the prediction are linearly related" (Vernoy & Kyle, 2003: 170). That was established through the correlations analyses. A multiple linear regression was indicated as a means of discovering additional factors that may act as predictors of academic success in addition to whatever predictive strength the CEPA might be assessed to possess.

3.6 Data analysis procedure

For the research study, the sources of quantitative data were the following:

1. CEPA English proficiency test results, which were taken before actual admission
2. End-of-Foundations-year final English and Maths results of the same cohort of students who took the questionnaire
3. Admissions data from the Ministry of Education and NAPO (numerical data)
4. Scanned and tabulated results of the closed-ended items of the college-administered questionnaire

The data gleaned from quantitative analysis was used to substantiate and support data collected and analysed from the student questionnaires and interviews. Much of the results of these analyses have been organised in tables and graphs in the following chapter for the purpose of greater clarity.

Another important element in the data collection phase was obtaining analysis records of the CEPA itself. This exam is regularly analysed by NAPO (National Admissions and Placement Office) for internal consistency (a form of reliability), as well as content and construct validity. Additional information about the validity and reliability of CEPA was provided in the responses about the content of these exams from the semi-structured interviews of administration personnel. Therefore, the discussion of CEPA results incorporated both quantitative and qualitative methods.

Developing Performance Indicators

An important, final outcome of this research study was the development of performance indicators based on the analysis of factors mentioned by staff and students as likely to have an effect on students' ability to succeed academically. One of the principle aims of the research was to investigate whether it is possible or advisable to make predictions about students' future performance based on their placement scores. Therefore, a range of possible factors were examined. These factors were primarily chosen on the basis of the results of the student questionnaire. Formulas were produced through the MLR which clearly revealed certain variables as good potential indicators of success in the college's first year programme.

3.7 Planning logistics (Site Selection and Access Negotiation)

The site of the main study for this research was a post-secondary, tertiary institution in the UAE which is government-subsidized and accepts for admission primarily only students recently graduated from secondary-level educational institutions in the UAE and other GCC countries. This institution was chosen for the facility of access and the researcher's familiarity with it and the staff since she was also employed there (EFL instructor for Foundations students). As has already been noted, written permission was obtained from the administration of the institution to conduct this research project.

3.8 Planning Techniques to Demonstrate the "Trustworthiness" of the Results (aka Validity and Reliability of Results)

Yin (2003: 38) described three kinds of validity that is necessary to establish for research: "construct, internal and external". Construct validity is based upon establishing that the research design follows a particular theoretical paradigm. Although complex and time-consuming, the construct validity of a case study can be achieved by a researcher if he or she is thorough and articulate about how judgements were made, which sources of data were used and why they were chosen, as well as sincerely attempting to identify areas potentially biased by either the researcher himself or the circumstances of the data collection. The paradigm of this study has been identified as a practical one (discussed at the beginning of this chapter). Also in this chapter is to be found the articulation of how judgements were made about the research sample and site, the data to be gathered, and how best to gather it. Issues of potential bias on the part of the researcher and the circumstances of the data collection process itself will be presented in the following chapter.

Patton (1990: 247) said that the credibility of research results can be strengthened through triangulation and he identified four different types. This research study employed two of them: methods triangulation and data triangulation. This corresponds to what Denzin (1988: 30) said about triangulation (also divided into four areas): using multiple methods and combining data sources. Within a MM study, especially a multi-phase, convergent MM study such as this one, the need to prove that the data collection methods were triangulated somehow seems rather redundant. Nevertheless, it remains a particular strength to be demonstrated, most often with subjective, qualitative data-gathering designs.

One of the intentions of this investigation was to investigate and identify other potential, contributing factors that may lead to academic success for these first-year college students by combining both qualitative and quantitative methods in the research design, covering both data collection and analysis, as much as was possible. Generisability in research often refers to the researcher's ability to make assumptions based on the study's findings (Lincoln & Guba, 1985: 290), meaning that the research design and methodology explain how the reliability and validity of the research was determined, which hopefully makes them trustworthy enough to generalise from. This also involves issues of causality (aka predictive validity), and is difficult to prove, i.e. that *x* leads to *y*. This is because the ability to determine causality is extremely complex. As many contributing, or even interfering, factors as possible need to be identified in order to justify the conclusions reached about why a particular phenomenon occurred. "In practice, three conditions must be met in order to conclude that X causes Y, directly or indirectly:

- X must precede Y
- Y must not occur when X does not occur
- Y must occur whenever X occurs

One way of determining the reliability of a research study lies in the ability of a separate researcher to be able to repeat the research, perhaps at a locale which is different but which has many of the same characteristics of the original study, and reach much the same conclusions. The best conditions for reliability to be established rely upon the researcher's modus operandi of his or her research being clear and well-organised. The conclusions arrived upon need to have as strong a base as possible in order to enable other researchers to concur with the results of the research.

3.9 Summary

The organisation of this chapter began with an explanation of the overriding paradigm focus of the study, as well as an explanation for the decision to mix methodologies. Thankfully, the "paradigm wars" of the 1980s and 90s have largely dissipated as scholars and academics have more and more acknowledged the complementariness and even compatibility of quantitative and qualitative methods that supports an array of MM designs. As Creswell and

Plano Clark (2011: 25-35) explained, the development and realisation of MM studies has gone through several important periods in the past 60 years. They call the present one the "reflective period", in which MM research is being critically assessed as a legitimate genre of research in its own right. It is hoped that this MM study contributes positively to the current methodological dialogue.

The decision to use a case study research method was explained in this chapter and the process of determining where and from whom the data would be collected was clarified. Then the actual data collection process was described along with the rationale for the decisions made. In the next chapter, the results of the data collection process will be presented in detail.

4: PRESENTATION OF THE DATA

4.1 Introduction

In this chapter, the results of the data collection process for this research study will be presented. The chapter is roughly organised according to the sequence in which the data was collected. Therefore, the findings of the responses to the student questionnaires are presented first. After that, the information gathered from the semi-structured interviews is summarised. The two separate group interviews with students is the final section of qualitative data discussed. This chapter then concludes with explanations of the statistical analyses performed – the quantitative element of this mixed method study, as well as the presentation of the results obtained from the different analyses.

4.2 Student Questionnaire Results

There were 37 questions in the student questionnaire. (The numbered items were 38, but as has been mentioned, number 26 was inadvertently left out of the numeration of the items. See Appendix B for a copy of the questionnaire.) Double class periods were organised (25 minutes x 2) as it was expected that it would take approximately 45 minutes per group to administer and complete (including explanations). It was administered to all the Foundations students in one day, 347 students, approximately 3 and 4 sections at a time. The researcher was to supervise the whole operation, but about 10 minutes before the first group arrived, she suffered a fairly severe fall at the college, and had to be whisked away in an ambulance. Thankfully, the Diploma Foundations (DF) and the Higher Diploma Foundations (HD) supervisors were well-advised of the planned procedures, and stepped in to take her place. Colleagues also assisted. The questionnaire was completed by the entire 'Foundations' cohort for the academic year 2007/8. The only exception was one HD section of approximately 20 students who did not participate because they were not brought to the auditorium to respond to the questionnaire. Additionally, one of the teachers who accompanied his section to the questionnaire administration advised his students not to write their names on the space provided. His particular instructions notwithstanding, all students were informed of their right not to write their names. The only difference was that the other students were encouraged to do so in the interest of supporting the research effort.

All parts of the questionnaire were presented in Arabic and English, side by side. As was mentioned in the previous chapter, the students recorded their answers on a bubble sheet (This was a generic answer sheet with rows of circles labeled a to d for each numbered item. Students darkened the circle which corresponded to the answer option they chose.). An overlooked typographical error resulted in the number 26 being left out of the question order. Even though the students were informed of the error during administration, still 25 filled in an answer for #26 (7%) of the 347 students who answered the questionnaire. These students were removed from the analysis of responses. In addition, while data were collected from the entire cohort, students who chose not to identify themselves (98 students, or 28%) had to be removed later (after tallying the totals) from the final group for analysis, since student names were required in order to include questionnaire responses in the Multiple Linear Regression analysis for the purpose of comparing different data about various assessment marks with their responses.

Part One of the Questionnaire

The first 17 questions (and numbers 25 and 30) were oriented more towards establishing a background profile of the students of this cohort. Several interesting facts revealed themselves as a result of the examination of the questionnaire results. For example, it was anticipated that very similar backgrounds would be reported, and this was indeed the case – for both levels – according to students' responses. Questions about gender and nationality (theirs and their fathers') were not included in the questionnaire because they weren't necessary, but they are included in Table 4.1 because of its focus on the background similarities of the students. For facility of reference, the questionnaire items included in Table 4.1 are the following:

#1: How old are you? a. <18 yrs old; b. 18-21 yrs old; c. 22-25 yrs old; d. >25 yrs old

#2 & 3: My father/mother ________. a. is illiterate; b. can read & write; c. is a high school graduate; d. is a university graduate or more
Optional - The specific level of education he/she reached: ______________________

#6: Where is your mother from? a. the UAE; b. a Gulf country; c. an Arab country other than the Gulf; d. She is not an Arab

#7, 8, 9: Which group contains jobs that are most similar to the primary one that your father/ your mother has (or had)/ you would like to have? a, b, c, or d.

#12: Do you have brothers and sisters? a. Yes, >10; b. Yes, >5; c. Yes, <5; d. No, I don't

#14: What is your marital status? a. single; b. married, no children; c. married, w/children; d. divorced or widowed

#16: Where will you live while at college? a. at home w/my family; b. in a flat by myself nearer to college; c. in a flat w/friends nearer to college; d. w/relatives nearer to college

#17: What kind of high school did you attend? a. public high school (government); b. model high school (government); c. private high school; d. home-schooled

Table 4.1: Students' Personal Details from Questionnaire Responses

Background Info	Questionnaire/Data Collection Results
Gender	100% male (Research was conducted at a men-only campus.)
Nationality of students	100% Emirati (Admission is only open to Emiratis & GCC nationalities. Even so, the students' admissions records did not reveal any other nationalities in this cohort.)
Age (questionnaire item #1)	91% between 17 and 21
Level of Father's Ed. (questionnaire item #2)	44% either illiterate or didn't finish school (10% & 34% respectively) (22% completed secondary; 34% university grads)[1]
Level of Mother's Ed. (questionnaire item #3)	64% either illiterate or didn't finish school (20% & 44% respectively) (21% completed secondary; 15% university grads)[2]
Nationality of parents (For the mothers, questionnaire item #6)	Fathers: 100% Emirati; Mothers: 80% Emirati, 17% other Arab nat.; 3 other non-Arab nat.
Occupation Boxes (questionnaire items #7, 8 & 9; adapted from a Yellis item, a test produced by CEM Centre, Durham University, UK)	**A** driver carpenter plumber electrician car mechanic hairdresser housewife **B** architect accountant doctor lawyer director engineer **C** civil servant pilot merchant teacher policeman politician nurse **D** farmer fisherman soldier security guard telephone operator computer technician medical technician
Father's Primary Occupation Mother's Primary Occupation	5% from Box A; 26% from Box B; 53% from Box C; 15% from Box D 75% from Box A; 6% from Box B; 15% from Box C; 3% from Box D
No. of Siblings (questionnaire item #12)	58% >5; 22% >10
Marital Status (questionnaire item #14)	94% single
Living situation while at college (questionnaire item #16)	96% with family
Type of Secondary School (questionnaire item #17)	88% graduated from public, government schools

[1] Interestingly, those admitted to HD reported much higher percentage of university grads (47%) than those in DF (27%). Conversely, a substantially higher number of fathers did not finish school amongst DF students (39%) compared to HD (26%).

[2] The educational level of the mothers mirrored that of the fathers in the difference between HD & DF students, only more so: 28% university grads for mothers of HD students to 9% for DF students. Similarly, non-completion of secondary was 33% for HD and 50% for DF.

Inclusion of questions about families' economic status might have ascertained more strongly that the demographic profile of the students in this cohort is as similar as it appeared to be, but this was strongly discouraged by the students who participated in the questionnaire pilot. Question number 9 asked students to choose the box of occupations they themselves were aiming for. The majority chose Box B (63%), with C (28%) as the second most chosen option.

In the optional clarification space for the questions about the educational level of parents (number 2 and 3), only a few students volunteered information about their fathers. This included 5 fathers who had Masters degrees (unspecified), and 5 fathers with a Bachelor of Science (one student identified this as a BS in Police Science and Detective Skills), as well as 3 more with unspecified university degrees. Unfortunately, the 4-option restriction allowed no option space for college diplomas, but six students volunteered additional information about their fathers. (This limitation is discussed in Chapter 6.) Most of the other clarifications were to specify how far fathers had progressed in primary, preparatory or secondary school. Two students pointed out that their fathers could read and write, but had had no formal education, one clarified that his father could read, but not write, and one student stated that he didn't know the answer to this question. As for the mothers, information voluntarily supplied included two mothers with unspecified Masters degrees and 8 mothers who had either completed a university degree, or who were in the process of doing so. As with the fathers, other information provided here was to specify how far their mothers had progressed in primary, preparatory and secondary schools. Only 2 students clarified an ability to read only, and not write. Finally, two responses in the open-ended final comments section referred specifically to these particular items themselves: (1) "Why does everyone want to know if my mother works? She does not work. She has never worked and she will never work, God willing!" (Researcher's note: He specified that his mother is a university graduate.), and (2) "I request that you not ask questions which are too personal (like the one about our mothers)".

Part Two of the Questionnaire

Questionnaire items 18-24, 28 and 29 were primarily about students' assessment experiences in the academic sphere. Bar chart 4.1 represents the results of items 18 to 21, which focused on secondary and CEPA scores, academic stream in secondary and parental pressure to excel. Again, for facility of reference, the questionnaire items included in Chart 4.1 are the following:

#18. What was your overall final average from your high school exams?
a. 90–100%; b. 75-89.9%; c. 50-74.9%; d. less than 50%

#19. What stream did you follow in high school?
a. Science stream; b. Arts stream; c. Started with Science and then switched to Arts

#20. Did your parents pressure you to study? a. Yes, always; b. Yes, but not so much that I felt overly pressured; c. Almost never; d. Only for certain subjects (Which subjects? ____)

#21. What was your CEPA English mark? a. more than 170; b. between 140 and 169.9; c. less than 140; d. I don't know. / I didn't sit for the CEPA.

#22. Did you take special, extra classes to prepare for the CEPA exam? a. Yes, in my school; b. Yes, outside my school; c. No: I did not think it was important or that I needed to; d. No: I wanted to but there weren't any extra classes available for me to take.

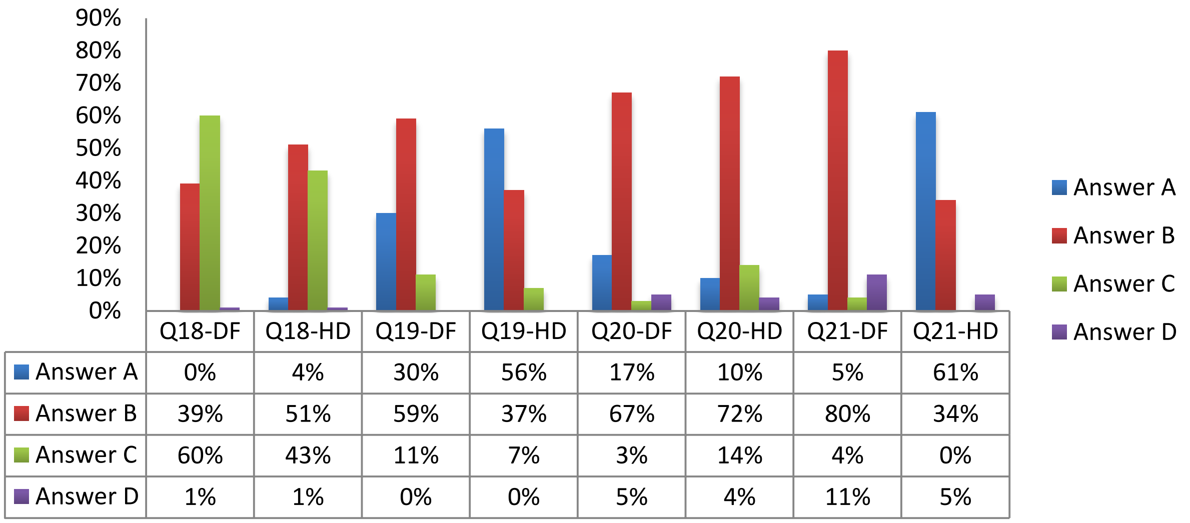
#23. How did you feel when you sat for your high school English exam? a. calm, relaxed and confident; b. a little nervous or worried; c. very nervous and worried; d. carefree

#24. Did you feel any differently when you sat for the CEPA English exam? a. Yes; b. No; c. About the same. If you answered yes, please explain why you felt differently for the CEPA

#28. Were you given instruction about how to fill in computer answer sheets for the CEPA exam by your high school English teacher? a. Yes; b. No; c. Not sure/ Don't remember

#29. Did your high school English teacher coach you about how to answer the kinds of questions you would find in the CEPA English exam? a. Yes. S/he spent >50% of our English lesson time doing this; b. Yes. S/he spent <50% of our English lesson time doing this; c. Yes. S/he gave us classes after school for this; d. No, not at all.

Chart 4.1: Responses to Qs 18-21, DF & HD



	Q18-DF	Q18-HD	Q19-DF	Q19-HD	Q20-DF	Q20-HD	Q21-DF	Q21-HD
🟦 Answer A	0%	4%	30%	56%	17%	10%	5%	61%
🟥 Answer B	39%	51%	59%	37%	67%	72%	80%	34%
🟩 Answer C	60%	43%	11%	7%	3%	14%	4%	0%
🟪 Answer D	1%	1%	0%	0%	5%	4%	11%	5%

Item number 18 asked students to report their overall final average (of all subjects) in the secondary final exams. (It is standard practice in the UAE to report a conglomerate average score rather than a GPA.) Generally, scores less than 50% are not acceptable for admission into any tertiary government institution in the UAE, but exceptions are made for extenuating/ mitigating circumstances. As would be expected, the majority of DF students reported averages in the range of 50-74.9% (item response "c"), whilst the majority of HD students reported averages in the range of 75-89.9% (item response "b"). There are two academic streams in public secondary schools: arts and science. Usually, higher scoring preparatory school students are placed into the science stream, which is considered more challenging. Therefore, it is perhaps not surprising that in item number 19 this trend continues into college, with the majority of HD students having completed science streams, and the majority of DF students having completed arts streams. For item number 20, both DF and HD students overwhelmingly affirmed that their parents did encourage them to study, "but not so much that [they] felt overly pressured" (item response "b"). At the suggestion of the students in the piloting of the questionnaire, a further space for clarification of item response "d" (only felt pressured to study certain subjects) was provided. Only 7 students (of the original 347) specified a subject: 5 said English and 2 said Maths. The responses for item number 21 also followed a predictable pattern. This asked students to report their CEPA scores (Item response options reflected cut-off scores for DF and HD at the college.).

Items 23 and 24 were about students' levels of anxiety sitting for their secondary English final exam and the CEPA. There was little difference between the results for the choice of item response "a" ("felt calm, relaxed and confident [sitting for secondary English final]") and item response "b" ("[felt] a little nervous or worried"). The results were 47% and 40% respectively. The responses for item number 24 ("Did you feel any differently when you sat for the CEPA English exam?") were mixed: Yes, 29%; No, 42% and About the same, 27%. The almost evenly matched responses were not particularly enlightening. However, Item #24 also included an open-ended response space which produced a large number of individual comments (60 responses). These comments were all written in Arabic (except for half of one comment). They were translated by the researcher (This translation was double-checked by two native speakers of Arabic at the college.) They were loosely organised into four major categories, with a fifth category for comments that were different from all the others. Understandably, since this was the first year that the CEPA became a high-stakes exam, 14 of the comments were about the fear and/or anxiety they felt about the CEPA. Feeding into students' natural exam anxiety was the fact that the CEPA was a completely different sort of examination than the examinations they were accustomed to. 19 of the 60 comments referred to this. At opposite ends of the spectrum, students expressed their preparedness for the CEPA (5 comments), or lack thereof (8 comments). Perhaps the most poignant group of comments centred around the awareness students had about the enormous importance of the CEPA result with regards to their future. Some of these comments may be found in Table 4.2 below. (All the comments to item number 24 can be found in Appendix D in the Student Questionnaire responses file.)

Table 4.2: Selected student comments from open response part of Questionnaire item #24

Category 1: Fear, Anxiety about CEPA

1. Worry and fear that I would fail in the exam and that I would not be able to go to college!
2. My friends told me that CEPA was hard, so I was a bit worried
3. I never felt such feelings before [about a test].

Category 2: Reactions to CEPA as a different sort of exam

1. I felt differently because the whole test was in English.
2. Difference in test atmosphere
3. The place was different. The number of people there [was different].
4. The format of the exam was so new to me.
5. It was a test at the end of the year. It was new for students. There was no chance to get accustomed to the level [of the CEPA exam] so that there would be no fear.

Category 3: Preparedness Issues and the CEPA

1. Well, English is the hardest subject for me because I am not good in it … I mean zero!
2. We did not know much about it and it was like the first time.
3. I didn't know anything about the CEPA exam.
4. It was so quiet and restful. There was enough time and what [we] covered had a role in helping me [finish] the test.
5. My English is not that bad and the CEPA was not that hard.

Category 4: CEPA and my future

1. I felt that I reached a new crossroad in my life.
2. Because it is an exam which determines my life's course. *Continued in English:* Because it clears the way for life if I pass, I will have the study in college or I will lose all my wishes.
3. This exam determines the course of my life, whether I enter a university, a college, or stay at home.
4. I didn't understand that my success in life would begin here.

Several issues are alluded to in these comments that require explanation, especially with regard to Category 2. The CEPA exam is entirely multiple choice except for an essay section, which is given a banded mark. The banded essay mark is explained in the interviews with the Foundations and CEPA supervisors. Students record their responses on bubble sheets. This was something most had never done before the CEPA. Also, the CEPA is a proficiency assessment, not an achievement test. Most Emirati students had never taken a proficiency test before the CEPA. The status quo to which they are accustomed is a system where cramming for exams and memorizing the materials in the syllabus to be tested with little synthesis of thought is the norm and the dependable way to achieve a range of good, or at least passing marks (O'Sullivan, 2008: 46). In addition, to facilitate the administration of the CEPA, testing centres were set up in central locations so that the entire UAE third year secondary student population could take the CEPA on the same day. For the first time, many students found themselves in a large exam hall with literally 100-300 students (or more) from several different schools.

It is important also to point out that the decision to change the status of the CEPA from a placement exam to a major part of admissions criteria for this nation's universities and colleges initiated a monumental shift in the public school assessment dynamic. It affected the de facto marginalisation of not only the GSC (General Secondary Certificate – awarded upon successful completion of the secondary school battery of exit exams) English final exam, but the entire GSC percentile score which had hitherto been used as a major part of admissions criteria. This enormous change was realised within the space of a few months and left most secondary English teachers frantically scrambling to not only complete the syllabus, but also to prepare their students for successful completion of the CEPA.

As reflected in the comments of Category 4, several students echoed sentiments of students around the world that the sum total of one's self and one's future resides in the results of one examination (or set of examinations) that will determine their future access to educational opportunities. Whether this feeling is justified will be taken up further in the following section about the interviews with college management officials.

Questionnaire items 28 and 29 queried students specifically about the instruction they received to prepare for the CEPA. Item 28 was suggested by an English teacher in a secondary school in the UAE. She was dismayed by how many students she noticed were unfamiliar with how to use bubble sheets to answer exam items during the CEPA. While 52% of the respondents did indicate that they'd been shown how to use the bubble sheets, it is of concern that 34% said they had not been, since this will have obviously had a negative effect on their ability to complete the CEPA as well as they might have had they known. Item 29 was in two parts and it was about the amount of coaching students received prior to the CEPA. The intention was not to ascertain if this actually affected scores, but merely exploratory – to include as many possible contributing factors to academic success for the multiple linear regression analysis. Of the 347 students who answered the questionnaire, 274 of them (79%) indicated that English lesson time was spent preparing for the CEPA. Only 3% said that after school classes were made available for this. The second part of item 29 asked students to specify the kind of CEPA preparation they'd received. Of those who answered, only 24% indicated that they'd been taught study skills, and only 28% were taught answering strategies. By far, the majority indicated that they'd been given sample questions and mock exams (55% and 42% respectively).

The final questionnaire item before the language acquisition motivation items was number 30 – one of the items that the students of the questionnaire pilot discussion strongly wanted included. 83% of the respondents answered the question: "Do you plan to work while you study this year?" This item had an open-ended response element which asked students to explain why they did or did not foresee difficulties with working whilst at college. This open-ended element generated more responses than any other item in the questionnaire (160). It should be noted that the college does not actually allow students to work and study at the same time during their English Readiness year. The comments of students not planning to work whilst at college, and those who were planning to work but expected difficulties were combined since many of the same issues were mentioned by both groups. These issues overlap, but basically include concerns about the lack of time to do both (25 comments), the excessive difficulty of doing both (51 comments), the distraction that work would be to doing well in one's studies (11 comments), and the pre-eminent importance of one's studies (12 comments) . A few students also pointed out that this was not allowed by the college (3 comments). Several students who declared an intention to work and study and who also expected difficulty with this, stated their desire to be self-reliant. 18 students altogether mentioned the importance of self-reliance. This comment theme was repeated in the final group (those planning to do both, but who did not foresee difficulty doing that). Students in this final group also expressed varying degrees of confidence in their ability to manage doing both. Some selected, illuminative comments are listed in Table 4.3 below. (All 160 comments to item number 30 can be found in Appendix D in the Student Questionnaire responses file.) Again, these comments were all written in Arabic (unless otherwise noted) and were translated by the researcher.

Table 4.3: Selected student comments from open response part of Questionnaire item #30

Part One: Students who anticipated problems with studying and working

Category 1: Time issues

1. Reviewing and studying lessons taken needs a lot of time. Therefore, there isn't enough [time] for a job.
2. Time is short and there is not enough time to review my lessons after I go home.
3. Time? Is there enough time to work AND study?

Category 2: Difficulty of doing both (working and studying)

1. If I start working, I will never complete my education.
2. It's impossible to do both work and studies well at the same time.
3. I'm afraid that I would not be able to manage things between study and work, or that having a job would be too taxing and so negatively affect my studies.
4. It's difficult for me to manage both of them.

Category 3: Work distracts from one's studies

1. [I won't work and study at the same time] because it will mean that I won't be able to devote myself to my studies.
2. It would take a person away from his studies.
3. Too much exhaustion and fatigue leads to a lack of concentration for my studies.
4. ... I don't care enough about my studies as it is.

Category 4: Study is the most important thing right now

1. Studying requires all of my time.
2. I will concentrate on my studies.
3. I just want to study, and I don't want to make things more difficult for me by working and studying [at the same time].
4. I will not think about working as long as I'm studying.

Other comments

1. I have a study leave. (from his employer)
2. Maybe because I never worked before, I don't know how it would be.

Part Two: Students who did not anticipate problems with studying and working

The need to be self-reliant

1. I want to work and study to depend on myself.
2. I will be depending on myself and taking care of my expenses until I graduate. Afterwards, I can undertake a profitable project.
3. I want to have experience in work before I graduate.
4. I want to depend on myself, build my own future and help my family.

Student confident in his ability to manage both well

1. I will be able to organise myself between work and study. I just need a simple job.
2. I have confidence in myself that I can manage myself between work and study, if I find [work].
3. I will specify a time for work and a time for studies.
4. If I work, I would be able to manage, but I think not working will make studying and doing well easier.

The final 8 questionnaire items (numbers 31-38) were language acquisition motivation questions. This addition to the questionnaire was suggested by Professor Vivian Cook, who also directed the researcher to use the shortened version of Gardner's 1985 motivation questionnaire, located online (http://homepage.ntlworld.com/vivian.c/SLA/MotTest.htm ; last accessed 6/30/11). The choice options had to be revised to a shortened Likert scale, since only 4 options were possible on the answer sheets that were provided gratis by the college. This meant that the neutral option of the typical Likert scale was not included. All of the students' responses to these questions were extremely positive, with 64-86% choosing the 'strongly agree' option choice, except for two of the questions: numbers 35 and 38. Students expressed much less interest in learning English in order to appreciate and understand English art and literature. Nevertheless, 53% strongly agreed and 34% agreed. There was even less interest in studying English in order to be more respected (41% strongly agreed and 33% agreed). This chapter focuses on the presentation of the results of the data collected. Discussion of the possible significance of the different results is the focus of the following chapter.

The last section of the student questionnaire was an open-ended one asking students for any other comments they may wish to make. Of the 347 students who answered the questionnaire, 59 commented (44 from DF and 15 from HD). 19 of these comments voiced thanks for the questionnaire. Many were pleased for the chance to express their views about topics of central concern to them, and for the researcher's interest in their opinions. Many of the other comments were suggestions and/or requests about college issues that bother them: most notably, the college policy about lateness (4 comments), the frustration students felt about the difficulty of transferring from DF to HD (4 comments), and establishing evening classes so that those who wish to can work while at college (5 comments). Several also requested changes in the way English is taught, including requests for enhanced remediation (6 comments). Five students made comments about things they thought the questionnaire should also have included. These are:

1. [My] answer to #16: I live in my own private home in Madinat Khalifa. [My] answer to question #17: I graduated from ADNOC Technical Institute.
2. Why haven't you included questions for students who are already working?
3. I think this questionnaire does not include that heavy information to declare the maximum things about the positive and negative of CEPA.
4. This questionnaire didn't ask about the reason for the lack of student interest in studying English.

5. There should have been a question about students in the DF programme whose marks are high enough to allow them a chance to transfer to the HD programme – marks which would enable them to take a test to switch to the higher programme. Thank you.

Finally, perhaps most revealingly, students expressed their hope that their responses – the results of the questionnaire – would be taken seriously: "I ask that those in positions of responsibility pay attention [to the results of the questionnaire]".

4.3 Semi-Structured Interview Results

Eight semi-structured interviews were conducted with different college and non-college mid-management personnel. These included the College Registrar, the college counselor, the supervisors of the Higher Diploma Foundations (HD) and Diploma Foundations (DF) programmes (2), the team leaders of the HD and DF programmes (2), a CEPA supervisor and the Director of Assessment at Academic Services in the Higher Colleges of Technology. All of the interviews were conducted between January and June of 2008. All of the interviews were recorded with signed, informed prior consent of the interviewee, and later transcribed. The content of these interviews, as well as the two group discussions with students are presented here. Some commentary is included here, but the summarization of significant points to emerge from the interviews will be specifically addressed in the following chapter.

Altogether, the total interview input was approximately 5 hours. This resulted in approximately 40 pages of single-spaced, transcribed interviews. One technological glitch resulted in the last 15 minutes of the interview with the Registrar not being recorded. The researcher relied upon her handwritten notes of his responses for these questions. The responses have been divided into college personnel's input (6 respondents), and non-college personnel's input (2 respondents). This data was further organised around the major topics of the interview protocol. In the interest of brevity and clarity, unclear syntactical utterances were edited, but the intent of the responses was preserved. In addition, interjected interviewer questions were removed whenever possible to preserve the flow of the responder's thoughts. Both of these changes may be noted in the use of brackets ([]), and the use of dots (...). The full, unedited transcripts may also be referred to in Appendix F.

4.3.1 Interviews with College Personnel

The following represents the key findings relevant to the research study resulting from the interviews with the college personnel involved in admissions, streaming, assessing and counseling Foundations Year students.

1: On what information are admissions decisions based?

The Registrar and both the Foundations supervisors were asked about this. All of those interviewed were aware that admissions decisions are based on a calculated formula called the "Placement Index", or PI, but only the Registrar knew how it is calculated. He said that it is a composite score comprised of the English secondary score (30%), the GSC score (16%) and the CEPA (54%, the two objective sections only), in addition to a separate minimum required score for the CEPA writing section. The result of this calculation is the PI. As the Registrar explained, "The PI last year for Diploma was a minimum of 29 and above. For Higher Diploma, a student needs a minimum of 71. A minimum of 71 to be considered, you know. Suppose you got a PI of 71, but still your writing band is too low? It doesn't work. You need a writing band for Higher Diploma of 3.5 and above." However, in spite of this, the HD Supervisor stated that she would only accept writing band scores of 4.0 (with rare exceptions).

Not unexpectedly, the Registrar was the most forthcoming interviewee with detailed information regarding the selection process for college admissions: "Ok, [the students] applied. They did the CEPA English and Maths. The Ministry of Education uploaded the results into their website. When the students have done CEPA English and Maths, they are eligible to be accepted. So, they [the MoE] open the site for us, for a week's time. Just one week. Researcher: "Higher Colleges is allowed access to the Ministry of Higher Education's website?" "Yes, only a few people have the admissions rights… only those who are dealing with admissions. They go into the site. They download the lists and their scores. When she downloads the scores, she puts them into [the PI] hierarchy.... I'll give you the full process: It depends on how many graduated from the Colleges [the previous academic year]. Central Services will give us a target number. Last year our target number was 1772. We look at how many students graduated and how many students are not returning, like failed, withdrew or whatever. We see how many are left. Let's say that after all we have counted, we have only 1200 students left. 1200, and our target is 1772, so that means we need to get 572 students. This is my target for admissions. But, if you say, 'I'm going to admit this number of students', well not all of them show up. It happens that after application and acceptance, people don't come. Our 'show rate' is 74%. So we also calculate this 74% so that the final admissions ends up as 572." Researcher: "A 'show rate' 74%? Isn't that a bit low?" "Always it's like that. But I'll tell you why. Because many times when they submit they also have other chances outside."

2: What specific criteria are used to place admitted students in HD or DF?

When the researcher asked the coordinator for HD Foundations about the role of CEPA scores in student placement, he described the CEPA as a "blunt instrument" and explained that all it does for them is decide who gets into HD or not. Interestingly, when this question was asked of the Registrar and the HD supervisor, they gave diametrically opposed responses. The HD Supervisor said: "HCT Central Services devised a formula… where the factors were more or less the same. CEPA English, High School English, CEPA writing band… and they came up with a range of numbers. And basically, they decided the cut-off point – this magical 71. So, any student who came out with a 71 or more with this formula, was allowed into Higher Diploma Foundations. Anyone with less than 71 went into Diploma Foundations. And, in a way, it took the decision-making away from the supervisor. Before, the supervisor used to scan the list and decide… and now it's sort of a 'done deed'. When I came back (from summer holiday), they'd already been divided (meaning placed into HD or DF levels). What I was able to do was look at a couple of borderline cases … a few students who had 70 point something, but who had good English CEPA scores, good writing band. But I still think that there may be one or two students who went into Diploma Foundations who could've come into our Foundations." She went on to clarify, "I suspect if I was here when the results come through towards the end of July, I'd probably still be able to control [decisions about admittance into HD]. So that maybe even if student who had 71 but who had [a poor mark] in Maths or something, then I could say 'Sorry, this student doesn't stand a chance of surviving the course'. Now really my role is more fine-tuning than making the bigger overall decisions because the big overall decisions [have already been made] made". Later, about this same question, the Registrar said: "We don't allocate. We say these are eligible to be in Higher Diploma or Diploma. We admit them, and then the supervisor of Higher Diploma will decide …'Ok, these students are good students for Higher Diploma. I'll take them. These students, for example, are not that good. I don't think they will do well.' So she will advise for them to go to Diploma. Then whoever's left in the list will go to Diploma because they are not eligible to go to Higher Diploma." The researcher is not sure how to explain this discrepancy except to put forward the suggestion that since the changes are new and many, not everyone is on board with who is doing what. For example, the HD supervisor did say that she'd been the one responsible for choosing students for placement in HD Foundations in the past, but since the change the previous year in CEPA's role, this was no longer her responsibility.

Some of the interviewees mentioned "challenging out" of Diploma Foundations. Theoretically, if a student is doing well in DF, and wishes to be given a chance to transfer to HD, he may take a "challenge test" for this purpose. If he does well, then he has a chance, at mid-year, to move up to HD Foundations. However, this does not happen often, as the HD supervisor explained: "And I say ... to students who say they want to continue to Higher Diploma. I say "We love to have you if you do the challenge exams and if we have a place, but you're not top priority." (Top priority being those who graduated from secondary that year.)

3: What percentage of students was accepted this year / is accepted?

In the interview with the supervisor for DF, the researcher asked her if the percentage of school leavers accepted for admission to the college was 100% or close to it. In response she said, "No, because we don't have the funds to open that many sections". She was asked what percentage did she think actually were accepted of secondary school leavers: "I would say nearly all of them this year got in. There were very few left alight". Researcher: "95%… 90%?" "Umm… I would say so. It'd be best to check with Academic Services. But even the ones who didn't get in possibly in September, in August, we're now having an intake of four new sections. So I would say… that we accept nearly everybody who ... Well no, cause some people want to go to HD, and they get don't get into HD, so they don't come at all. [They're accepted … but they decide not to come] … " To this, the Registrar differed saying, "We give a chance to those who did not have a chance before – the students of this academic year. And that's why they say we're taking 100%. We are not taking 100%. And not even 100% will show up because many of them are not interested." Taking into consideration students accepted into the spring semester, the Registrar said, "It would be then about 85%... 85% of those who apply are offered a place in the College." Like the DF and HD supervisors, he also cited budgetary restraints as being the most important limiting factor in the number of students offered a place at the college.

4: Marks weighting scheme for the college foundations year

Questions 4 and 5 were answered only by the DF and HD supervisors and co-ordinators/ team leaders. As for the weighting schemes, all agreed that the only fixed system-wide weighting framework was 70% for course work and 30% for the final exams. "There's of course an outline and there are outcomes. And you have to meet those outcomes, and you've got to cover what's in those outcomes, but how you do it is up to you" (excerpt from interview with DF supervisor).

5: Who is responsible for setting HD/DF weighting schemes? ... and for the assessments that comprise the major component of the final mark?

Within the course work mark, the system allows supervisors carte blanche to decide how their particular department in their particular college branch (There are 14 branches of this college in the UAE.) determines the elements of the course work, and how each of them will be weighted within the course work mark. At the time of the interviews, a major change had been initiated two years earlier in the weighting of the course work for HD Foundations at this particular college. The HD supervisor explained: "It changed a couple of years ago because it used to be skills based – you know – reading, writing, speaking, listening. It used to be a straight even split: 25, 25, 25, 25 [respectively]. Now, for the past two years, reading has been the most heavily weighted at 35%, writing is 30%, listening is 25% and speaking is 10%. That's how it's weighted now." Course work is weighted differently depending on the college branch. The HC supervisor described how they break down course work in Higher Diploma Foundations: "The course work has different elements. It has progress tests throughout the year – 7 progress tests – for the course work. Then they also have a reading mark based on reading tests throughout the year. They have a writing mark for writing done during the year. They have a careers mark. They've got a library skills/research skills mark". She also pointed out how this freedom to proceed within the course work mark affects teachers: "And this filters down to the classroom level too. I mean, basically teachers can do what they like in the classroom as long as their students are ready for the exams. They can use the books that are recommended; they can [decide] *not* use the books". Researcher: "Interesting. So, they basically have learning outcomes that they have to guide students towards and however they get there is up to them?" "Yes. There's a course outline with learning outcomes, but [we've been teaching this course] for quite a while under a very good team leader who has designed most of the materials and he's built in all the learning outcomes into the materials. Very few people actually, you know, bother to look at the course outlines because they know that using the materials, everything's covered."

The DF supervisor was at the time fairly new in her position. This may explain the confusion in her response to questions about the course work mark and assessments: "Yes, [the course work elements are] obviously different according to the semester. The first semester is worth 30%. The second semester is worth 70%." (Researcher's Note: In the calculation of the students' final marks at the end of this one year foundations course, 70% of this aggregate 100% course work mark is taken in order to accommodate their final exam scores, which account for 30% of their final marks.) "I suppose I decide, with the team leader, or co-ordinator, what the weighting should be. It can be [different for every college], because some colleges do projects, for example, I mean I think what we actually teach is up to us, but most colleges do the same. We have a curriculum leader who sets certain guidelines. So yes, we could decide to be completely different, but generally speaking, I would think we all do the same thing, because we all have one curriculum leader".

Perhaps the most interesting responses to this particular query were the ones about how student assessments are determined and prepared. Both the HD and the DF supervisors relied on their respective co-ordinators to prepare periodic written assessments for students in both levels, and they revealed very different viewpoints with regards to materials preparation. In the interview, the researcher asked the HD Foundations co-ordinator what his role is in assessment decisions and organising assessments. At first, he replied that he had no role in such decisions until he understood that the question was referring to the periodic tests which are part of the course work mark, called 'progress tests': "Oh! Progress tests. Oh yes! Sorry, sorry, I thought you were talking about CEPA, because at one stage we used to get people from the schools with the results from the schools and we used to take no notice whatsoever of these results and administer our own exam. We're not allowed to do that now. We have to accept the students that we're given. Then all the progress tests… well, I wrote them and we administer them". Researcher: "Ok, you write them? What's the process of developing a progress test?" "Um, well they mirror the final exam completely." Researcher: "Is this the June exam or the January exam?" "It's a year course so it's the June exam. We familiarise students with the… we teach to the test. So we familiarise them 'til they're bored stiff with the format of the exam. We don't do anything else apart from using exactly the same order as it is in the exam. We've used the same exams for about 8 years.

We polished them. So we know which ones are difficult and which ones… And they work." Researcher: "So you have different versions?" "No, no. We keep control of them and oddly enough, the results are, year by year about the same." It's interesting to note here that the DF supervisor pointed out that HD Foundations at this college had the best pass rates in the whole system (of the 14 branches). She went on to explain what she believes is the explanation for this: "... the reason they have the best pass rate is because, of course, they can select who they take in, and anyone who doesn't perform is moved down to the Diploma Foundations in semester two [in January]. So they [retain] their favourites".

When asked what was her role in assessment decisions and organising assessments, the DF co-ordinator explained her process: "Assessments are all in my hands. Having said that, when I took over as Team Leader and then Coordinator, I was told very clearly that it's 30% KCA (Key Common Assessments = final, centrally-prepared examinations) and 70% coursework, but … I was told that here at [this college], the supervisor expected there to be formal assessments on each skill, each semester on which we base our coursework. So it would be easily defended should it be questioned. I was given that guideline, but in terms of setting when the assessments are and writing the actual assessments, the tests, that's all in my hands". Researcher: "How do you write the tests?" "I have a bank of some tests that were used in the past. I like to encourage group participation [of teachers], but very few participate. ... I mean, as a teacher myself, I don't get paid any extra to do it. … We know what we're teaching. We know what we're supposed to teach, and it would be nice to have a say. It shouldn't be just one person (developing these assessments), because maybe my approach is not the best approach. And maybe that's not how most people are going at it. So, I think the motivation for them would be nothing to them personally, but perhaps professionally: making sure that their programme was well-rounded, and that they had some say, and that they knew that what they are teaching was going to be assessed properly. But there's no release time, or anything financial or anything like that." Researcher: "Do you have different versions of the same test?" Yes. Listening's a bit more challenging [to have multiple versions of]."

6: Is the college's reputation in general a factor in terms of 'cost'?

The college's student counselor, Registrar and the supervisors responded to this question, however the counselor requested that his response to this question not be included in the transcript of his interview. In general, the feeling amongst all was that this was not a major factor in admissions decisions. The DF supervisor pointed out the importance of reaching enrolment targets for funding and maintaining an acceptable pass rate: "Unfortunately, we have to take in the greatest number because the college needs so many students to get funding, to get whatever it is they get from whoever it is that gives it to them, ... Government funding basically, and if we don't meet our targets… We needed 1700 students to get our funding, ok, so we took into Diploma Foundations more than 300 students. Ok, so yes your point - we took them all in, but they weren't particularly good. We don't have any choice about who we take in because we have to take in the right number of students to get funding, and that funding is for all the other programmes." In fact, the major issue that seemed to concern the DF supervisor with regards to the college's reputation was how displeased the college director would be about a poor pass rate.

The supervisor of Higher Diploma maintained that financial considerations were much more important than their exam scores: "It's funding. Funding is what really determines what we've got". When asked this question about whether the college's reputation was a factor in admissions decisions, the Registrar said: "Yes, it is a factor, but it's not a factor that will stop us from getting students in. It's a factor in *placing* students. If a student is not up to the level of Higher Diploma or Bachelors, we don't take them into Bachelors. We take them into Diploma. Basically, if a student is not at that level, he will be placed into Diploma. For sure, every student will have a level to excel in. Diploma Foundations is a very basic level of English, Math and other courses. We take them into this level... Anyone who's able to cope with study, they continue. ... Our standards remain the same. If we take them, and they can't do well, they drop out. But with this strategy, we are giving a chance to everybody. They can't say: 'We did not have a chance'."

7: Is any potential added burden to college staff a factor in admissions decisions?
Matching the economical response of the HD supervisor to the previous questions, when asked this question, the DF supervisor responded, "No, not at all, not at all. It's money." The HD supervisor provided a candid reply: "It's not such a big factor in Higher Diploma Foundations because we're getting the top end of students anyway. We do stream the students. We end up with some low level groups. Teachers are asked what kind of students they prefer to teach, and some of them actually prefer to teach the weaker students."

The interviewee who really elaborated on this point was the DF co-ordinator. To begin with, she said that supervisors were more pressured than teachers to achieve higher scores. She continued, saying, "Obviously, it makes teaching a bit more challenging. You have a set of goals and objectives that you're supposed to meet and if they're not coming at the level we anticipated when those goals and objectives were created it makes it really difficult. I don't think teachers feel the pressure, but there is a lot of pressure from the powers that be on supervisors. ... They get the pressure to increase scores. We were told very clearly that it was expected that at this campus that our pass rate go up by a certain percentage. Great. Easy to say, but when the students come in lower than what we expect them to come in with, all of them making progress, there's no way… It's very unlikely that they're going to. We're lucky to have them meet the exit criteria, [let alone] exceed. I think that's a big burden, and we feel the heat. We feel pressured to push the students at a faster pace – we as teachers – at a faster pace than the students are perhaps ready for or that they can handle".

8: That which may hinder students' ability to succeed

Borderline acceptance and possible psychological and emotional effects on 'at risk' students

When the HD Foundations co-ordinator was asked about this, he minimised this particular type of effect: "No, I don't think [it has that kind of effect]. They are very, sort of, results oriented. If they've been told that they passed, I think they're very happy, whether it's a good pass or a bad pass, doesn't matter – it's just a pass. Researcher: "If [borderline students] pass and they make it into this system, do you think that they struggle?" "Oh they do! No doubt about it. Does this have a negative effect emotionally or psychologically? No."

The HD Supervisor didn't go as far to downplay this effect: "Yes, of course. Like students [anywhere], sure there's a bit of extra pressure there. Worry about what's going to happen next year. I mean, I've had students come to me, 'Oh, I failed my reading or I failed my listening. Am I going to Diploma?' And I've got to reassure them that there's no question of that... That we're looking at the whole semester. And then next semester, the weaker ones start to worry if they'll pass at the end of the year and what that will mean and is there anywhere to go [for help]. Well, there is. [She referred to the ALZ: Active Learning Zone: a college centre for student remediation] We've also got quite a few sponsored students and of course, they worry about what their sponsors will think if they're not performing well, So

I think it's no more than for any group of students anywhere: the psychological worry of succeeding and failing."

As for Diploma Foundations, the DF Co-ordinator was very specific in her opinions about this kind of effect on students: "I think it's really demotivating when the teacher's throwing a lot of stuff at you and you're not able to grasp it and the teacher says, 'I'm sorry. We've spent all the time that we can on this. We have to keep moving forward.' That's got to be demotivating and frustrating. … There's gate-keeping at 1255. It's right before the PET [Preliminary English Test, produced in the UK]. There's a test they take, like a 'pre-PET'. If the teacher doesn't feel that … you're going to score high enough on the PET, you won't get the opportunity to take it. You can officially pass the course, but they will not allow you to move on to the next level." Researcher: " 'If the teacher feels…?' Isn't that a recipe for misuse of the system?" "Umhm… (agreeing sound) So I think it really harms our students coming in really low. Even if we push them through, they still don't have the command that they should have. Then they take 3 or 4 months' break where they speak no English [during the summer holiday], and come back, obviously having lost half of it. Then [they have] only a semester, maybe two maximum, to try to get it all back and progress to sit the PET. I know some students who did really well the first two years and then get [stuck in] this holding pattern. They've 'finished' all their coursework, but they can't sit the PET because of their English. And you can't graduate until you pass the PET. Other colleges have gate-keeping, but they do it much earlier. [It's probably better] because then the students, if they're given the opportunity … to repeat the course, [this] gives them more time. ... Some colleges just automatically dismiss them, which I don't agree with either." [Researcher's Note: The college no longer had this requirement after 2008.]

The DF supervisor echoed the sentiments of the DF co-ordinator: "[I]t's not nice to be in a situation where you're failing. I mean they might be better if they were put in a situation or a context where they could excel, whatever area that might be - not academic, but something else. But to constantly be put in a situation where they're going to fail. I think it's very demoralizing. Yeah, it must have an effect. [I deal with this sort of thing] quite a lot". Researcher: "How do you manage that kind of problem?" "Well, how do you manage it? I mean a lot of it, to be honest, has to be done through the counselor because the students, generally speaking, do not communicate well in English … I have lots of time for students who just cannot do it. What I don't have time for are the students who have potential and don't use it. So, the ones who are just struggling but who are trying - you know we send them to the ALZ [Active Learning Zone: a college centre for student remediation] and we offer them extra help. We offer them counseling. We tell them about the other options – but there aren't many other options for them, although I think [the counselor] might know more about that than I do.… How do we help them? It's through counseling and listening to them just giving them, you know, time."

Perhaps not surprisingly, the college counselor had more to say about this than any other interviewee, and he spoke at length about the difficulties many borderline, or at-risk, students face at college in terms of psychological and emotional effects, settling in difficulties and other sources of struggle. After acknowledging that he did not have empirical evidence for his opinions and that he was generalising, about their emotional reaction to college he said that most entered college with "happiness and hope". He pointed out that most of the students are aware that their educational experience thus far has been poor, that they are aware of their lower level in college subjects than others, but that most view their college acceptance as a good opportunity. Perhaps obviously, he cautioned that individual attitudes had much to do with a student's ability to improve and compete with other students. He said that those who have a positive attitude "take advantage of being here and they quickly bridge the gap between them and the other students and they quickly grasp the concepts and learn and respect the rules of the college, and learn to be in a college and do their best as far as learning. So they end up kind of working on their level and they bring themselves up academically. Some of them they fail to do that, for a variety of reasons. Now, the ones who fail for a variety of reasons… maybe they fail because they have personal issues at home, maybe they fail because they have a horrible poverty problem or transportation issues, or maybe they fail because they have special needs problems - learning problems, visual problems, psychological problems, depression, anxiety, and so forth… addiction problems, smoking, internet addiction, and so forth. Those are the ones that become very, very negative. And they become very hard to deal with in class and on campus. Those are the ones who end up creating a lot of problems to themselves, other students, faculty and staff."

Settling in difficulties

The DF supervisor was very much aware that "settling in" difficulties could negatively affect students' ability to excel, especially in the first semester of the term. To this query, she responded: "[This] is why semester one [course work] is only worth 30%. We look at the first semester as learning how to get here on time, and how to behave and how to relate to the other students and property and then the real work starts in semester two, and I think we do see a big difference in them between one semester and the other". Interestingly, the HD supervisor considered this less of a problem for her higher level students, and also she felt that it takes them much less time: "Difficult to quantify that. I think it does take them a week or two to settle in, to get used to their teachers. Again, I think a lot of that will depend on what kind of background they're coming from because some of them are coming from government schools outside of Abu Dhabi. This would be a huge change for them. They're in a mixed gender environment and English is all around, when they've probably not heard that much before. But we're getting increasing numbers of students coming from private schools where English has already been a medium of instruction. So I think it would depend on their backgrounds. But I must say that they're generally very adaptable students who seem to settle in ok. I mean, we've got very experienced teachers too. They're tuned in to these things; they've been here quite a long time. They can tune in to the comfort level of students and do their best to help.

The HD Foundations co-ordinator felt that encouraging a sense of superiority over those who didn't make it into HD smoothed over the settling in process: "[This does not affect them negatively] because we make a point of congratulating them that they have succeeded in getting into the HD Foundations course and we sort of make sure that they feel good about this. So, in fact they feel that they are successes". Like the HD supervisor and co-ordinator, the DF co-ordinator also placed importance on the ability of the teachers to facilitate students' settling in to college even while maintaining that, in her opinion, settling in was not an issue. She explained that the real issue was the process of retraining them as students: "It's not really settling in that's the issue. It's just that we're going back to step one. It's kind of like me trying to sell them my ways. [I say], 'I know it's different. I know you don't understand it. Let me explain it to you. Let's try.' And really, doing it again and again and again until they see that it does work. They're used to a teacher talking to them 100% of the time and telling them to memorize these words and these grammar points and throw papers and fill it in. My way is more … Pretty much the whole class is in their hands. I have a lot of pair and group work which is something they're not used to… I have a lot of card-based activities which they're not used to. Frankly, they're settled, but they're not quite sure it's going to work. And they're skeptical. I'd be skeptical too. So it takes time to build them up to where they realise that they can do it on their own".

For his part, the college counselor viewed this as a serious issue, and named the huge difference in teaching strategies and the educational experience at college in general as 'academic cultural shock'. He pointed out that this is primarily an issue during their first year at college. He said, "I do believe that it's hard on most of them, [but] it's easier on some. And I do believe that most of them end up dealing with it within maybe a year... the first academic year – the foundations year in Diploma and Higher Diploma… It seems to me so far it [is] the year in which they work – the good students – they work mostly on that cultural kind of front. It's just a personal observation because it seems to me that by the Diploma 1 or Higher Diploma 1 [after the foundations year], most of the students ... seem to be settled in. They seem to be more in harmony with themselves, other students, staff and faculty… more in harmony with the college rules … more in harmony with the whole environment."

Other sources of struggle

The interviewees were asked to volunteer any other factors they could think of which they deemed a source of struggle and hindered students from excelling. When the college Registrar was asked about this, he singled out "[t]hose who are academically weak, meaning those who have poor study skills [and] those who are unwilling to ask for or seek help, [perhaps from] shyness". Apart from the difficulty of having a low level of English, the HD supervisor also pointed out that "[s]ome of the students have an awful lot of commitments outside – just tons of problems for them. Especially maybe the ones who are married and have a family. Some of them have business interests. Some of them, even though they're not married and don't have their own family, they may be the man of the house because their father's died or something, so they take the whole responsibility. They end up having to do a lot of driving and other family responsibilities. So all these things impinge on their ability, even if they're quite motivated. I've also come across a few students who've had a real confidence problem as well. They're able, I can see that, but they don't really believe they're able, and they take an awful lot of encouragement". The DF supervisor suggested that "maybe they should be advised early on that they don't have an academic bent, and they should possibly look somewhere else – possibly to Work Readiness". The HD Foundations co-ordinator felt that students struggle because they don't apply themselves: "The standard is not high. We have a year, and if we had students who worked *averagely* hard, we could probably do the work in four months. We can take *anybody* through … get them through the exams that they have to pass. They're not difficult. We haven't got any EFL students coz they've all been filtered out by the CEPA, so the material we get is quite adequate and we have about a 90% pass rate. The other 10% could pass, but they waste their time".

The college counselor identified three other factors that he felt had a major impact on a student's ability to succeed: attitude, excessive distractions and the misconceptions many students have of being in college. He enumerated them as follows: "If they are negative. If they have some sort of low self-esteem. If they end up in the wrong crowd. Sometimes a positive, enthusiastic student ends up in a wrong crowd of students who are negative, failing, feeling like they are losers, feeling like they have no hope. Those are things that affect negatively the students' opportunities to succeed. There are other reasons. There are a lot of factors involved in answering this question. When students end up getting addicted to the Internet or chitchatting, or starting smoking, or start to drive around the city and waste time. And the students get this misconception also, which is: 'College is a place to have fun.' That's a misconception. 'College is a place of freedom. I can do anything I want.' That's a misconception. 'College is a place where I can do anything I want, the way I want. Nobody is watching over me because I'm not at school anymore.' That is a misconception. 'College is easy.' Misconception. 'College is just a place where you go and you attend and then you'll pass by yourself.' That's a big misconception. Another misconception sometimes I get from students is 'Oh, our teachers don't like us. They are going to fail us. Our teachers are horrible in this class. Oh, my teacher doesn't like me. She doesn't like my name. She doesn't like the way I look. She doesn't like the way I walk. Oh, she hates me and she's going to fail me'. That's a big misconception. I never came across one teacher here who really hated a student and she or he failed him. That's a misconception for some of the students that might prevent them from really succeeding".

9: That which may help students to succeed academically

Even though the counselor labeled as a misconception some students' views that they are failing because a teacher has taken a dislike to them, he mentioned the role of the teacher as a key factor that could help students: "As far as the teachers go, I think some factors that will help students succeed [is] if the student has a teacher who is devoted, a really devoted teacher; a teacher who doesn't suffer from misconceptions or pre-conceived notions. I really believe that it [could] destroy the students' opportunities if the teacher of the students has [misconceptions] like, 'Oh those students are really academically weak and they have no chance'. When the teacher gives up on the students, the students I think are going to have a very hard time succeeding. It's usually when the teacher is optimistic, when the teacher is devoted, when the teacher believes in the students that the students will be able to make it. I think that's an enormous, wonderful opportunity for the students to succeed".

The counselor also mentioned factors "relating to the student himself": "I do believe that students who are positive will end up succeeding more than students who are negative in general. Students who have a higher self-esteem, I would say that this will help them to succeed more than the rest. Because being positive and having a higher self-esteem ends up helping the student to deal with the issues they have, whether they have transportation issues, poverty issues, special needs issues… so they end up dealing with their issues. [Also] when the family is supportive, optimistic, positive, encouraging, understanding, flexible… usually that ends up supporting the student in succeeding in the college".

10: Students' motivation/s to learn English

All the interviewees strongly indicated that, in their opinion, the strongest motivation to learn English was to get a good job. "Good job" means one that pays well and is in a field preferred by young Emirati men. This would be those careers included in box B (architect, accountant, lawyer, director, engineer) of item #9 in the student questionnaire, which was chosen by 63% of the student respondents. Their responses follow:

- HD Foundations co-ordinator: "Umm.. probably primarily a job. They are aware that English is a world language and that they have to function in English. On the other hand, most of them can already function adequately for holidays, et cetera, so for the majority, it's sort of an academic box that they have to put a check in. I don't think that they have any particular interest in English per se."
- HD supervisor: "To get a good job! Without a doubt. They want to get a good job, to earn enough money and that's it.[Nothing] about English language itself. But because it's an international language... They recognise it as an international language. They can see it used around them all the time. They know if they're going to study Engineering, or Business or IT they're going to have to use it. They know they need it, but they're whole motivation – main reason – for studying, not just English, studying anything, is to get a good job."
- DF co-ordinator: "Based on what they tell me, my students have said that the number one thing is a job. They all want to get a job, a decent-paying job when they finish. And

obviously, in this country, that does many times require a good command of English. Occasionally some students tell me, 'Well, you know, English is the language of the world, so we need it for everything'. But number one reason for them I believe is a job."

- DF supervisor: "Most of them, to find a job. I don't think I've ever met anyone who's doing it for the knowledge of the English language. It's just to find employment - or some of them because their parents have told them to come here."
- College Registrar: "Many have a weak background in English. They wish to communicate with others. They need English to get a good job."
- Assessment Co-ordinator, Academic Services: "Umm… external, external motivation basically… to enter ... the HCT environment, to enter a credential programme, to achieve the credentials to graduate. Perhaps for some of them to get a job. So I think it's basically external [motivation]."

11: Students' attitude/s towards learning English

The college's Registrar mentioned "having a good attitude towards studying" as a factor that would help students to succeed in general, but when asked about students' attitudes towards learning English, he said students "think that English and their poor marks in English, or their poor ability in English prevents them from getting better marks or results overall". The HD Foundations co-ordinator alluded to this problem as well, but in a different light: "They know how practical it is, so we don't have to sell the course. They have friends, they see them, and if that friend can't speak English that person is effectively… um… Well, completely marginalised. He can't do well". The HD supervisor believed that students in general have a very positive attitude towards learning English, and she made another valid point about the country's leaders: "I think [the students do have a positive attitude] - well, in Higher Diploma Foundations, they do. If you ask them to write about it, it's always in very positive terms. I think they really do see it as an international language... Their leaders within their own communities see it as such. Certainly Sheikh Nahayan sees it as an absolute essential for any development of the country and so on. I think that message gets through to them as well". The DF supervisor echoed the same belief that students generally have a positive attitude towards learning English, and she added, "Most of them see it as a way ahead".

The DF co-ordinator provided a revealing perspective on the attitudes of students learning English at the college: "I think from the conversations I've had with students, they seem to be very open-minded to learning English, however, I do think that they are quite a bit jaded against learning English, based on their history of learning it throughout their education. ... I mean, these boys have been studying English for years and years and years and have not even, well, a basic command at best. So I think that's very frustrating. And I've had students say to me, 'Well, look at us. Ten years of English and we speak like this. What can you do for us?' And I've had them ask me, 'Why is it that in France, Germany and Spain.. Why do they speak English better than us? You teacher, when did you start learning French? You had less years learning French than we did learning English. Why do you speak it well?' I think they want to learn, but I think their educational background, or their experience in learning it has really kind of programmed them against thinking that they can acquire it and master it without a lifetime of study."

12: Other information from the interviews

Both co-ordinators were asked whether the CEPA results generally match teachers' perceptions of what their students' abilities are. The HD co-ordinator stated that "we never pay attention to that. Once we get a class list, that's our class list". The DF co-ordinator criticised the results as being inflated from coaching, an observation also commented upon in the two non-college interviews. She began by comparing last year's with this year's CEPA scores: "I personally looked at the CEPA scores for my class this year and my class last year, and on paper the CEPA scores for my class this year are much higher than last year's. So, initially I thought, 'Oh! I'm going to get a really solid group this year.' The writing bands were this year around 2. I think the CEPAs were maybe 140, but not exactly sure. But they were definitely higher than last year's. So I thought I'd tweak my lessons. … and I was in for a big shock when I got my students. Because in fact, the CEPA scores were higher, but their actual abilities in the English language skills were much lower than last year's… with students who had lower CEPA scores. [The students] told me [that they] were prepped and coached. Essentially given the test… again and again and again. They were given mock exams for the writing. They were given mock written paragraphs, letters, and they memorised them all… on every topic one could think of. They were coached. They were prepped. They were paid an injustice in my opinion, because they're coming in with these scores, and we're thinking, 'They can do this'. If you prep me for any test, I might do really well, but do I know it? Can I use it? Do I understand it? No! It's easier to coach CEPA than IELTS or TOEFL. It's not this big, huge, vast area that anything could be pulled from. It's very confined and restricted. That's my opinion."

The Registrar stated that the P.I. generally matches what students' abilities are, but mostly for English (as a subject). He also mentioned that students may not ever re-sit the CEPA. This means that, unlike the TOEFL and IELTS, there is no opportunity for any student to take the CEPA a second or third time. At the data collection stage of this study, the CEPA was a one-shot examination. (This policy changed to allow multiple sittings of the exam in 2011.)

The HD supervisor ended the interview with a query: "I'd like to know from Central Services what CEPA score they use [to calculate that predictive validity score]. Do they just use the objective score, just the Reading and Grammar, or do they tie in the writing band as well? (Researcher's note: The PI did not include the writing band score at this time.) I have a feeling – just an intuitive feeling – that maybe we shouldn't place so much on the writing band because of the subjective nature of the marking. I have re-marked some CEPA writing... where you know, they've sent me up scores for students joining late. They've got very good CEPA scores generally and maybe a 3 or 2.5 in writing. So, I've asked for the writing scripts to be sent up and when I re-marked them, I might be a good band out from what they were originally marked. I just worry about the writing banding and wonder if the predictive nature of CEPA could be based solely on Grammar and Reading. That would be fabulous because it takes out the subjective part.

And finally, the college counselor made an interesting observation about the whole institutional system of the country for admittance to post-secondary education. He pointed out that "some of our students are [academically] poor. And the factors…why are some of our students poor? If we look at the academic structure of the UAE, we will [note] that [this college], Al Ain University and Zayed University are the federal governmental [choices for tertiary] education for UAE nationals. Zayed is for women [only] and Al Ain University, the United Arab Emirates University in Al Ain, is also for nationals, but it's kind of structured in a way that the United Arab Emirates University will take students who ... have a high school percentage of above 70%, I think. They have the right to study in the United Arab Emirates University. Students who score somewhere between 80 and 90, they will get the chance to get a scholarship and go abroad. Students who score above 90%, they will definitely take the opportunity of studying in the States. So it seems that students who are very academically high achievers, they get the chance to go abroad, right? And less will end up in the United Arab Emirates University… and the bottom of the group end up in [this college].

... I don’t know if teachers know that or not but that’s a fact of life here. The best students don’t come here. They go to the States. Why [would] a student who has the right to get paid to be in the United States – getting a degree from the United States – [come] to [this college]? As far as I know, when I did the recruiting last year, almost none would be interested in this option. Almost all of them would end up in the States. And I used to see them in the States. Very high academically-achieving UAE students. The bright UAE students end up in the States. You won’t see them at [this college].”

4.3.2 Interviews with College-Affiliated Personnel

The last two interviews were conducted with managerial officials directly involved in assessment development, one for the central administration office for all fourteen branches of the college, and one for the CEPA itself. They were queried about several issues, some of which touched on issues covered with the previous group, and some uniquely connected with their specific responsibilities and knowledge.

The Assessment Co-ordinator from the college’s Academic Services Department

The Assessment Co-ordinator was asked about his role in assessment decision and organisation. “I provide advice about reliability, validity, practicality … and the impact of what we do in terms of assessments. We’re constantly looking to try to change the mindset in a way, of seeing the exams not as a student exit system, but as a way of measuring progress as they move through the system – the idea of exams, and giving various feedback to colleges on areas to concentrate on. We want to retain students, to bring them through. That’s what we do: we educate. We should not be getting rid of students because that’s how the testing ways are done”.

The focus of his position is the internal maintenance and improvement of the college’s assessment tools. Therefore he did not have much to contribute in the way of information regarding the admissions process or the CEPA or the Placement Index. Even so, he responded to several of the questions posed to the college interviewees. He concurred with others who stated that the reputation of the college itself was not a factor for admissions decisions. He did qualify that however with regards to admission to what he termed a ‘credential’ programme, i.e. the Bachelors programme. For this, he said, the college required students to achieve an academic IELTS band 6, which he described as an internationally recognised benchmark of language competency at least (if not more).

He also provided information regarding the KCA examinations, which are the system-wide, summary assessments for the end of the Foundations courses (HD/070 and DF/0155). He said, "I actually inherited this - how the calculations are made about [the way] they arrive at a pass mark. It's a Policy Council decision. There are four papers in the final exams: reading, writing, listening and speaking [for both 070 and 0155]. ... The papers are weighted exactly the same. The weighting for reading is 35%, for writing is 30%, for listening is 25%, and the speaking is 10%. They're exactly the same, 070 and 0155. And students are required to achieve an overall mark of 60%… an overall pass mark of 60%. That's the exam. They can fail any number of papers providing they make it to 60%. Now, this is based upon an analysis done by my predecessors. The two of them sent it to the policy council in 2004/2005. I've not seen the research paper. The weighting is there. It's to do with predicting academic success … 'post-Foundations' at both levels. I don't know where this [analysis] is. I've been trying to find this for quite some time. Researcher: "Why can't you find it?" I have no hard copy nor soft copy available here. Policy Council papers are not stored in an accessible way, let's say. I was actually… just before you came in … that's exactly what I was doing. This is it. It says that … I'll read to you what I've …[He peruses his monitor screen.] (Reading) 'It's an analysis of academic factors contributing to the successful student progression at HCT. This was requested by the then Vice Chancellor and will be presented at the forthcoming Policy Council. In a nutshell, the key points … that the skills of English could be aggregated if they are weighted. All previous studies indicated the success of this policy.' It's to be presented. It's a study to be presented. Ah.. They have no record of the study. How odd. [The policy paper] was done by [a predecessor] ... who left before I arrived. I've been here nearly two years (in 2007). So she prepared the paper and it went to Policy Council. ... And so they approved this weighting, this aggregation. So the [idea] was that the aggregation would remove the need for a re-sit. Researcher: "There are no more re-sits?" "No, no. That was the intention then. There are now because it's considered a 'must pass' exam and it's felt that it's unethical for the students not to give someone a second chance on a must-pass test. It decides their future academic career, you know. So we re-introduced the supplemental (the re-sit)."

As pointed out in other interviews, the periodic, summative assessments given to Foundations students throughout the term – Progress Tests – are also not system-wide, but determined by each individual college. This was affirmed by this co-ordinator: "Progress tests are college-based. They're not sort of centrally-produced, administered or marked, so colleges design their own instruments according to their different populations. [We in Academic Services do not] have any role in that process. This is in the domain of the individual colleges. See, strictly speaking, there are two divisions. There are system-wide assessments, which we're responsible for… They are the KCAs… Now these can either be must-pass or not must-pass. College-based assessments which are in the hands of the colleges … they cannot be must-pass. They can only be a part of the assessment scheme."

The Assessment Co-ordinator clearly defined the different programmes available at the colleges within the discussion of the different requirements for them: "in order to get into a Bachelors programme, they need an IELTS band 6. Then to Higher Diploma, which is a credential, a sort of sub-degree credential, then that's academic IELTS band 5. ... For Diploma and Diploma Foundations, this is sub, sub-degree by some way. These are certificate-level courses. Diploma Foundations in the past has been very open." At this point, he echoed what the Registrar spoke about in regards to the kinds of opportunities the colleges can potentially offer: "I think that the role of the [tertiary] institution within the UAE which is more to provide further education for UAE nationals [than to achieve an international reputation]. There are many who have weak language skills. The issue we've got is how we can best address those students' language skills to get them up to the level where they can benefit from ... You know, this concept of ability to benefit. Can they benefit by studying through English? Now, we can move them through the Foundations Programme to that level of English where they can benefit by studying in English, and we can move them through certificate programmes. What we want them to do is to progress through the Diploma and move to Higher Diploma and on in the future hopefully to the Bachelors. We don't think it's going to be thousands, but the opportunity should be there for someone who's motivated, who's prepared to work and take responsibility for his own learning to move through." But there's that important caveat: for the student who's motivated and can take responsibility for his own learning.

In a different part of the interview (noted below), he stated that English competency is a critical issue in institutions where the medium of instruction is English, but he added that "obviously there's far more to it, isn't there? There's motivation, which is a key thing. There are lots of things in the affective domain which really are sort of a kernel in someone's success or failure." He commented on several factors which he saw as hindering a student's ability to do well:

- "[A CEPA score of 140 or less upon entry] isn't a burden – it's a problem. They've got a lot of work to do. If you come in with a very low level of English proficiency, to get yourself up to the required standard to benefit from studying through English obviously requires a lot more work on your part… improvement? Whatever word we might use in there… They've got a ways to go.
- "Again, we have no empirical evidence, but experience tells us that [settling in difficulties] will have [a negative effect] on some but not on all of them. My understanding of college life [here] is that they are quite good at trying to bed students down quite quickly. I know that it's variable across the system. Some colleges are putting great efforts into student support, in particular identifying students at risk – for whatever reason .. on evidence or on reports from teachers and dealing with them early on. And there are reports of success.
- "The message that I often get from colleges is that those students who aren't succeeding – the ones 'at risk' – have a problem dealing with the idea of being in a learning environment where they also need to take responsibility for their own learning. They want to be spoon-fed, coddled. 'Teacher, I don't understand. The teacher's not very good. I didn't do very well.' So there's no sense of responsibility."

Conversely, he elaborated on motivational factors that might positively affect a student:

- "[Students' motivation for learning English] is an external motivation basically…I would think to enter, sort of, within the HCT environment, to enter a credential programme, to achieve the credentials to graduate. Perhaps for some of them to get a job.
- "[Things] that facilitate a student's success at this college [would] certainly [include] their level of proficiency upon entry. [Also,] their motivation for being here – Whether they've chosen to come because they want to pursue an academic degree, or whether they've been 'parked' here for some time [just waiting until] they're accepted into police college or the military … because that's what they do."

About what he would consider a 'clear accept', the Assessment Co-ordinator spoke of the value of the CEPA and said that he thought a CEPA score of 165 would be "a good indicator of future success of the [HD Foundations] KCA". He continued to expound upon this topic, touching on the need for more research and the value of the KCA assessments and the P.I. He said, "Now to what extent the KCA is a predictor of future academic success on the credential programme … We're in need of a longitudinal study to look at that, because we don't have that yet. This is everybody's next project because we need to know. ... This is what the Placement Index is driven by, if you think about it. They're looking to see how this – if you like – basket of scores is a good predictor of academic success. And they will play with it… They say: Alright, it may well be that this score, this raw score, that comes out … we see a Placement Index of 61, may represent different figures in the future. You know… different scores on the test because as a predictor, it's seen to be effective or ineffective. It seems to be quite effective. I mean the scores you've quoted me were much lower than what I had recalled, but still way above international predictive validity studies which range between 20 and 30. So this is still very, very promising. Yes, it is. For me the fundamental issue is the curriculum for Diploma and Higher Diploma. You need language – you need English because it's an English-medium institution. If your language skills aren't there, it's not to do with your cognitive ability nor your academic ability, you just can't access the curriculum."

Even though the Assessment Co-ordinator stressed the importance of high correlations between the CEPA and the KCA examinations, he also provided a seemingly contradictory explanation as to why it is difficult to make straight comparative correlations between them: "The CEPA standardized score is based only on vocabulary and reading. There's no assessment for speaking. There's no assessment for listening. The writing assessment is a little restricted. That's what we look at in the KCA scores, particularly the 070 (HD) scores. So we look at these scores, and we say to what extent are we looking at black and white, because our 070 score is a composite, quite a range of language skills… The CEPA standard is very narrow in its focus... It does not cover as much as the KCA. As a really useful instrument, it's a bit flawed. We are not convinced that it's the best way of providing positive information to colleges to enable them to change what they do – to improve the student learning. CEPA is an admission tool. Once we've got the students, we have to have a measure which brings them through. CEPA doesn't provide a measure of the skills. It provides us a score for admissions purposes. But what *we* want to see is what improvement the student made from when he came to us and at this point here. How can we feed that information back to colleges so that they can improve what they're doing in teaching and learning processes? That's what we're looking for. [The writing band does assist in this function] in a way, but in the past it's been a limited task; it's been a letter or something. … Right now, we've looked at coaching letter writing – which isn't itself representative of academic writing. It's very formulaic and you can memorise chunks to write. When you look underneath the skin, [you find] that it's not very good evidence, and not doing what we want it to do. We want it to feed back to inform our processes." It would appear that while everyone wants to have a predictor of future academic success, and acknowledge that CEPA has achieved high correlations, they also acknowledge that there are difficulties in understanding why these high correlations happen because of the difference in instruments and skills measured.

He was asked about his opinion regarding the change of CEPA to a high-stakes assessment and whether the CEPA had been better as a diagnostic instrument used for college placement (as it had been prior to 2006). He responded by focusing on the fundamental difference in purpose: "Different tool, not better. Different purposes. I think it's reasonable for the UAE to have such an instrument like CEPA for admissions – it's reasonable. As an instrument, it's being refined. People are responding to the data that's getting back, the information received from colleges coming back to them about how students progress. I think it's all fed back in. It's an instrument that's developing. I don't have an issue with it. I think it's reasonable there is this instrument that's being used in this way. I think it's reasonable the way it's being developed. It's changing, responding to observations. [It doesn't feed back into our system because] ... it's for a different purpose. The purpose is paramount in using an assessment and what we're looking for are assessments that feed back into our students' learning. *What the CEPA is providing is a measure of an ability to benefit in study.* We look to see its predictive value. We feed back to NAPO about that and say 'Look, this is how it's going'. It's useful data for that." Researcher: "But isn't just the fact that it's part of admissions criteria mean that you're using it to predict future academic performance?" "Yes, because we have nothing else. … I think it's ok [for its purpose]."

At the end of the interview, the Assessment Co-ordinator summed up by saying, "I think the message you're getting from me is that what we are looking [for] at HCT is the way we look at English assessments; not so much as gate-keeping instruments, [but] as informing the system. What we want is to facilitate learning, to improve their English. CEPA is an admissions tool. It's useful in terms of admissions and students' future progress. Its purpose for me is complete on admission." It is somewhat incongruent for the Assessment Co-ordinator to make this statement when he himself is in charge of posting the correlations of CEPA marks to final English marks at the end of the Foundations year for all the 14 campuses of the college. Obviously, the CEPA's purpose for the college is not complete on admission.

The CEPA Supervisor

The supervisor spoke at length, and in detail, about the CEPA. The interview was divided into two main parts: first, "Students and CEPA" and then "The CEPA Itself". In part one, he stated that not only do 75% of students who take this test score above the minimum criteria, but also that this is a big improvement from five or six years ago, when only 50% would have accomplished the same achievement. When asked about what actually was the minimum requirement, he said, "Well, it depends on the institution at this point. The Ministry of Higher Education three years ago set 150 as the minimum score for entering tertiary study, but different institutions are beginning to choose their own minimum scores. So, UAEU will probably choose a higher score – maybe a 155 or 160 for entry into the university - that's for the Foundations programme. They will not be considered for admission at UAE University unless they have a minimum score of 155 or 160. I don't think the decision is official at this point, but they are certainly considering it. And the other institutions are as well." Later on in the interview, the supervisor also added that the function or role of the CEPA is not to replace other benchmark exams, because – as the Assessment Co-ordinator mentioned – it's purpose is different: "[I]n fact, [all the national, tertiary institutions in the UAE] don't accept the CEPA score in 'lieu of'. UAEU will use the CEPA score to identify students who can directly 'challenge', but they still have to present a TOEFL or IELTS score. The CEPA fulfils a very important function because it allows institutions to place students in different levels of instruction in the Foundations levels. ... Right, now at the level we're using it – 150 … TOEFL and IELTS don't discriminate at that level. They're useless for any purposes below … really below 170 on the CEPA. So the CEPA scale starts much lower than the TOEFL and IELTS and doesn't go nearly as high as the TOEFL and IELTS. In the current context, CEPA cannot be replaced by TOEFL or IELTS. They will not provide the necessary information to these tertiary institutions."

In commenting on factors which could contribute to a student's success, he first touched on what a score of 150 in the CEPA means: "Well, the 150 is actually a very low ball. I mean it's a soft pitch. Most of the students who score 150 are eligible to enter the tertiary institutions, but the students who are at that level will … almost all of them will fail … will flunk out … will not be successful." He continued, explaining what would probably be the implications of a CEPA mark of 150: "Really, the fact of the matter is, that if a student scores that low, they don't have very good study skills, across other subjects as well, and tend to be at-risk because of that. Obviously, students at the 150 level have a low vocabulary, a poor command of grammar, very, very limited writing skills, and really are not able to function meaningfully in the language. So, in order for a student to get beyond that, there needs to be some engagement – as a student – with English. That's really what we're seeing." He was also asked to comment about factors that may help students do well in the CEPA. He immediately differentiated between public and private education: "Well, with private and public schools there is a huge difference. Private school students tend to do very, very well on CEPA. It's the public education kids who make up most of the students who aren't successful in CEPA. There used to be a strong difference between rural and urban students, but that seems to be breaking down a little bit more as there is increased awareness out in the more remote areas that there are standards that students are going to have to meet. Students are really stepping up to the plate and performing." The CEPA supervisor also described efforts by the National Assessment and Planning Office (NAPO) to facilitate an even better pass rate, called the Professional Development Instructors' (PDI) Programme. NAPO employs ten instructors to work in different educational zones to raise the awareness of supervisors and teachers about more effective teaching methods. They are also responsible for locally orienting supervisors and teachers to the expectations of the CEPA. The CEPA supervisor described the PDI as a very active programme. In spite of this programme, the CEPA supervisor also pointed out that the PDI is supporting a general, country-wide trend towards improving English language competency.

In the next section, the results of two group discussions with select groups of students are presented. One of the issues that arose was some students' belief that ever since non-UAE nationals started taking the CEPA, the test became more difficult. This issue was thoughtfully addressed by the CEPA supervisor: "Many students perceive it to be more difficult, but it isn't. What it is, is higher stakes. Once you make it an entry requirement for tertiary study, [that means] that there's an actual 'cut' score… Before it was an entry requirement, you had to take it, but that was all. You're already admitted. Now you're just being placed. So, it was a requirement for placement, but not a requirement for admission – but everybody had to take it. But once you say that everybody has to take it, *and* you have to get a minimum of 150, or else you can't come to our institution, suddenly the test may *feel* more difficult. In fact, the [students'] perception might be right. Suddenly, it's more difficult. Is it more difficult to simply take a test, or is it more difficult to attain a certain score? It's more difficult to jump over a line than it is to walk over a line? Maybe they're right, but the actual difficulty of the test has not changed".

The final comment of part one from the supervisor was about his opinions regarding students' attitudes towards learning English. Admitting that this was something NAPO had not explored yet, he continued by saying that it "has changed quite a bit over the years". His opinion is that "students who are studying English right now think of English as an international language which gives them access to higher education, and then also, something that they see their parents using it as well". He added, "I think we're getting into the second generation of kids where many of the parents also use English on a daily basis as well. This is either to manage their businesses or to go to Starbucks and order a cup of coffee. It's something that is so apparent to everybody that you very rarely see truly negative attitudes towards acquiring another language".

As was mentioned, Part Two of the interview focused on the CEPA itself. To begin with, the supervisor explained that the score system of the CEPA is a 'scaled score'. The supervisor explained that "all scaled scores are theoretically unbound by zero and infinity. So the CEPA, as one of those 'scaled scores' is unbound at the higher end." Arguably, one of the most interesting responses began as a response to the researcher's question about the CEPA's validity. The supervisor began by clarifying the construct validity of the CEPA: "[T]here's a couple of different validations. One of them is, of course, the theoretical validation through the description of the construct – construct validation. [We have achieved that] through looking at the different item types that we've brought to bear on the problem, and how those relate to theories of language. For example – you know – Bachman's framework. We can see the grammatical competence and vocabulary and different things and how those relate. The kinds of tasks that students are given for writing and how we expect that to conform to theoretical ideas about how [proficient] student writing is done. The [item] specs were developed with [these theories in mind]."

Then the supervisor proceeded to comment on two other types of validation: concurrent and predictive. He said that they've undertaken studies at NAPO where the same students sit for the CEPA, the TOEFL and the IELTS examinations. He said, "We would expect high ability students in one test to be high ability students in the other test and vice versa. And that's the case: [CEPA] does have strong correlations with both the TOEFL and the IELTS". As for its predictive validity, he began by noting that the CEPA is a "predictor of ability", mentioning the high correlations reported. However, his comments then shifted to the kind of language abilities that may be revealed by these examinations: "Generally speaking, the students who are entering university with acceptable TOEFL scores or IELTS scores have attained a level of English that is sufficient for them to actually begin to do things in their classes. And language becomes less of a factor and other factors start to kick in". The CEPA supervisor previously stated that the CEPA was designed to discriminate at a lower level than IELTS or TOEFL, and that it has demonstrated its usefulness for that purpose, but when queried about CEPA's correlations for scores of less than 170, he revealed that scores below 140 indicated that students simply did not – for whatever reason – engage in the test. "The lower-end scores, there's a lot of other things going on. Once you get down below 140, there's a lot of random guessing… A lot of that is a failure for the student to engage with the test to begin with. I call it the "yella effect". (Researcher's note: "yella" is an Arabic word meaning "C'mon", as in "C'mon – Let's go!) It's a case of [arbitrarily filling in all the answer bubbles as quickly as he can] and off he goes. You end up getting students who come in with, let's say, a 50... Nobody gets a "50", alright? It's kind of a fictitious score." Of CEPA mid-range scores, the CEPA supervisor said, "The strong correlation with future academic performance is probably because a student is moving into an English medium environment at this level: 150, 160, 170. Language is such a huge factor for them. And 150s simply do not have the necessary language skills. At 160, they're *barely* capable of learning in English. It's not until 170 that you really begin to see students who have both the study skills and the academic orientation to make up for their inadequacies and to really begin to learn whatever their content subject is - in English. That's why we have such a strong correlation [at 170] and you don't tend to see that with the other studies that are being done around the world because those students tend to be so much higher anyway. I mean, no university in the United States or in England would ever accept a student with a CEPA 150 or 160. They're just too low."

The supervisor was asked about the reliability of the CEPA as a correct estimate of students' English ability. In his response, he began by reiterating the high correlations achieved by CEPA and students' first year final marks, and then he spoke about the CEPA's internal consistency: "If it wasn't reliable, you would see low internal consistency in the test itself. We have very high internal consistency. At the item level and at the student level, the scoring is very consistent. **...** We have the item-level information on the test data. A classical test statistic would be like the KR20 or the Cronbach-Alpha which measures the extent to which you have inter-item correlations. Typically in high-stakes tests, you want something that is over .90. We're consistently at .95 or .96 in all of the tests. We've *never* been below .94. And that's as good as anything the TOEFL ever gets, and they have 140 items in their test. We have 120 items. Still, we're absolutely at a par with them with internal consistency. And then you know, also anecdotally, it's very rare that a student is misplaced into a level at UAE University or HCT or ZU. You just don't see it very often. But we *do* see it, and there's almost always an explanation for that. Some students do successfully cheat. We do see cases where a student who has practically no language ability will be placed in level 3. It happens once every couple of years. You'll have one student and they will say, "How did this happen?" Well, this student cheated successfully. But it doesn't happen very often and the fact that this is a rare occurrence and that people notice is also strong evidence of the reliability of the test."

The supervisor was asked whether he thought that the homogeneity of the student population contributed to these high correlations recorded: "Possibly. I think that, for the most part, the item types that we have on the test are not the kinds that you usually see having a differential function given different backgrounds, for example Japanese or Chinese or whatever. Once you move into speaking tests or listening tests, you do tend to see some pretty dramatic variety with mixed-background students. But when you have students from a single background, the real homogeneity of the group that we have here, we wouldn't expect to see any changes … So, I don't think that's really a factor because we are just simply focusing so strongly on vocab., grammar, reading… The writing may have some background, but that's not factored into the [scaled] score that they use for entry into the university." Continuing his response, he also connected this with the reliability of the test "I think it's very highly reliable, and it's also a testament to the reliability of the instruments that the institutions are using for their mid-terms and finals. If the UAEU mid-term or final exam had low reliability, then we wouldn't have a good correlation. So, you have two tests that are functioning very well in the environment that they've been designed for. So then the correlation shows this. Any correlation you have is going to be attenuated for the reliability of the exam. The mid-terms and finals produced by HCT and UAEU and ZU tend to have very high internal consistency as well. So, they're measuring language ability. Ours is measuring language ability. And they match up very well. CEPA is one half of the coin. Tertiary institutions are making some very good exams as well."

Earlier in the interview, the supervisor mentioned comparisons to IELTS and TOEFL scores. He was asked about a table produced by a presenter at a local conference in which he compared CEPA to IELTS and TOEFL scores (see Table 5.1 in Ghazali, 2008; He was a secondary English teacher in a UAE government school at the time. Reproduced here as Table 4.4.). This query revealed efforts by NAPO to establish equivalencies: "Well, what's he's done is … Well, we do have information on the IELTS and TOEFL that is … I don't know if it's publicly available, but it's certainly something that we've sent out in many reports. The Ministry of Education has it. It's probably posted someplace. So those equivalencies are probably pretty accurate." Researcher: "How did you arrive at them? There are different skills being tested in the three different exams, correct?"

Table 4.4: CEPA-English Scores vs. IELTS and TOEFL (Al Ghazali, 2008: 14)

TOEFL Paper	TOEFL Computer	TOEFL IBT	IELTS Equivalent	CEPA-English
625-680	**263-300**	**113-120**	**7.5-9.0**	**220-240**
600	**250**	**100**	**7.0**	**211**
575	**232**	**90-91**	**6.5**	**202**
550	**213**	**79-80**	**6.0**	**194**
525	**196**	**69-70**	**5.5**	**185**
500	**173**	**59-60**	**5.0**	**176**
475	**152**	**49-50**	**4.5**	**167**
450	**133**	**39-40**	**4.0**	**158**
425	**113**	**29-30**	**3.5**	**150**
Less than 425	**Less than 113**	**Less than 29**	**Less than 3.5**	**Less than 150**

"Sure - … anytime you do equivalencies, you look at the average performance of students in a given band. If, for example, we had a thousand students take the IELTS exam shortly after taking the CEPA exam… You'd look at each of the IELTS bands and say there's 150 students in Band 5.5, for example, and then what is the average CEPA performance for those students in Band 5.5? That's how you establish the equivalencies for those. We do the same thing for TOEFL. That's what we did. You bet! That's how we arrived at the equivalencies for those… We've had more than that, actually now, over the years. It's been something that's gone on for quite some time because all of the institutions do require IELTS and accept TOEFL. We have ended up having quite a few students who have both of those scores. And actually, we sponsored a study about 18 months ago through UAE University, then across the country as well, where students – after taking the CEPA – were then given the TOEFL and the IELTS exams and we produced this chart. Now, what they've done here is they've taken the three scores that we do report, which are the IELTS, the pen-based TOEFL and the CEPA, and then they've extrapolated from that for the other three columns. So the TOEFL IBT, they just linked that up to the regular TOEFL. And the CEPA English for expatriates – That's the Ministry of Education score. That's something that he must have extrapolated from his students. When you say that 150 students contributed to this, that's probably where he generated that from. The IBT correlation is readily available on the ETS website. That's where this came from. But we didn't produce this table. But it looks about right, about what we would expect..."

Finally, the supervisor explained NAPO's rationale for not including a listening or speaking element in the CEPA as essentially one of logistics: "We've got 35,000 students across the country, the test needs to be done on a single day, a typical interview will be about 20 minutes, it would have to be a one-on-one with a trained examiner. … for us to do it in a single day, we're going to need 3000 examiners, working across the country, on a full day – and that doesn't count the support network involved. It's logistically impossible given the current constraints. If they relax the constraints of having it all done in a single day, still it would be very challenging logistically, and it's one that the institutions themselves have not felt necessary to add in. I think that it's important to measure that, but we're just simply not at a point where we can do that. Computer-based technology may allow us to move in that direction in the future, but we're not there yet, either. That's the reason for the speaking. For the listening, listening also has a logistics element. One is that we administer the test to more than 50 venues across the country in two different shifts on a single day. You'd have to calibrate 50 different sound systems in different rooms and manage the distribution of the tapes, or whatever you're going to use. Acoustics vary quite considerably depending on the venue. So there's an issue of fairness. There's also the problem that you can only have one version going at one time. Right now we have 3, sometimes 4, versions being administered at any shift. So, a student sitting next to another student – they're going to be on different tests.

Whereas if you have a listening exam, everybody's on the same page, at the same question at exactly the same time. The cheating factor just goes through the roof. It tends to be a part of most language tests that is most amenable to cheating and also most sensitive to external influences. Given those considerations, and the importance that the test be highly reliable for these institutions for the purpose of placement, we didn't include listening. That doesn't mean that we've ruled it out for the future, it just means that at this point, and due to logistical considerations, it wasn't included in the original test design."

4.4 Group Interviews

Two separate group interviews were held at two different times in the academic year with mostly the same students from those who responded to the questionnaire and for whom there were CEPA scores. A decision had to be made about which of the 300+ students would be asked to join this discussion. To begin with, a fairly arbitrary decision was made: to start with the students who'd not only identified themselves on the questionnaire, but also were vocal enough to write a thoughtful or relevant comment at the end. The researcher reasoned that such students would be the most articulate and forthcoming in a group interview format. The first interview was conducted a month after the administration of the questionnaire. The second was near the end of their academic year, as they approached the time for their final exams. This was done in order to gauge any differences in their opinions, after having successfully navigated their course to the end. The full transcript of these discussions may be found in Appendix G. (Two students who participated in the first group interview had dropped out by the second. They were both contacted by phone. One did not respond and was reportedly suffering personal problems; the second did respond and did attend the second group interview. He dropped out because he'd found employment that he considered was more suited to his needs than a degree.) All of these students were individually contacted, either in person or by phone, and asked to participate. Approximately half of those agreed (27 students), and of this group, only 13 students actually took part in the group interview. A question framework was prepared in order to provide direction to the discussion, and was based directly upon anomalies uncovered through the collation of the student questionnaire results of the research study, and the collation of the questionnaire responses (both closed- and open-ended). The majority of questions in the group interviews were open-ended in an attempt to collect undirected responses. "Questions that include words such as how, why, under what conditions, and similar probes suggest to respondents that the researcher is interested in complexity and facilitating discussion" (Stewart and Shamsadani, 1990, p. 65).

For both groups, they expressed their preference to converse in Arabic. Their responses were recorded, with their permission, translated and transcribed by the researcher in order to facilitate the analysis of the discussion. The goal of this analysis was to single out strongly-held majority opinions, as well as to identify patterns that recurred with both groups.

Several questions were directly inspired from the sub-questions of the research questions: B.4. What are their attitudes towards learning English? and A.2. What is the predictive validity of this English proficiency exam in gauging the ability of students to progress successfully in English courses?

4.4.1 The First Discussion: December, 2007

Length of Discussion: 90 minutes; Number of students attending: 13 (all male)

The discussion began by asking the students about what their personal reasons, goals and/or motivations are for learning in English. One student said that they need English for their future, and many corroborated his opinion. He said that we are in the age of technology, and English is the recognised language of technology. They also mentioned that English is the lingua franca in all the airports and banks. Another added that even here in their own country, officers in the Army must learn and use English. They mentioned that it is the only language through which they can converse with people from all nationalities. Another student said that in the UAE, you can't get much done if you can't communicate in English.

Other responses:

"If you speak only Arabic, almost no one will be able to understand you."

"It is a good thing to know another language." "Why?" "It makes you a cultured person. It is even a tradition of the Prophet Mohammed."

"What kind of job would you find if you could not speak English?" "You could be a private in the Army. Take an ordinary position in the police, maybe."

"We need English for any college in the UAE. English is the language of study for all the college and universities in the UAE."

At this, one student said that he'd spent three months over the summer in the UK to learn English. He said that he'd learned more worthwhile, practical English during that time than in all the previous years of his education. And another thing, he said that studying at any tertiary institution in the UAE will not enable a person to learn English well. He said that he'd forgotten much of the English he'd learned whilst in the UK.

The second question asked was: "What is your attitude towards learning English?" This got a very positive response. They all feel and understand the need to be strong in English in order to communicate effectively with non-Arabs, and in order to get a good job. The responses overlapped from the previous question, many reiterating what they'd said in response to the previous question.

They were then asked how they feel about education and learning in general. There were some very interesting responses. They were generally proud of their government's initiatives to ensure the education of all its nationals. They spoke about their parents' generation – how back then, it was easy to find a good job with just a secondary diploma, or even finishing the preparatory level ; that illiteracy was very high then and anyone with a bit of education could get a good job. They said that this has dramatically changed. Now, so they said, it would be very difficult to get a decent job with only a secondary degree. One student mentioned the story of his own brother, who is studying in the U.S. His uncle told him not to return without a Masters. They seemed to have mixed feelings about this trend. On the one hand, it was of course, good for them and for their country that so many are now educated, but on the other hand, they expressed concerns about where it will lead. One student said, our children may not be able to find a good position without a PhD. "How can everyone have a PhD?"

The next question was: "What kind of support, encouragement, and/or pressure did you / do you get from your parents? Some cited a lot of pressure for certain subjects, especially English and Maths. Others said that they were under heavy pressure to study during secondary school. Another mentioned that his parents would intensify their efforts to make him study before exams. Two students said that most of the input they received from their parents was in the form of advice and encouragement. Then they were asked how many of them were now living with their parents and siblings. All of them said they were, except for one student who was living with his uncle and his uncle's family. I asked if the input was the same now that they are in college. They unanimously responded that the kind of support they were getting now from their families was advice and encouragement, and not heavy-handed pressure. In the questionnaire, most of the students indicated that they come from fairly large families. Almost 60% indicated having 5 or more siblings. Their opinions were canvassed as to whether they thought this had any effect on their ability to study or succeed, and if so, is this a positive or a negative effect? It was expected that some might complain about interference from younger family members or the difficulty of enjoying a good study atmosphere in which to concentrate, but their responses were totally positive. They cited the atmosphere of competition generated amongst brothers and that this was very motivating. A couple of others mentioned the desire for the children of their brothers and sisters to look up to their uncle – that this desire encouraged them to work hard to be successful.

They were reminded that several of them had mentioned the need for more attention to English language study before the college level. They were queried about the problems. One student pointed out that he felt that there were several unnecessary subjects, like Geology and Geography, that they had to take in secondary school which could be removed in favour of more English classes. They spoke about Arab teachers teaching English and almost never using English in the class. They said that these teachers translated everything into Arabic and even taught English grammar in Arabic. For these reasons, they never took English classes seriously. It was quite a shock to have to give English such emphasis during their last year of secondary. Most felt that the government should employ native speakers to teach English, but interestingly, not at the primary level. They felt that students at that age were too 'vulnerable' to have a foreign, non-Arab teacher.

Does the quality of education differ between schools outside of Abu Dhabi city and those in the city? Perhaps predictably, students who studied in schools outside the capital said that their education was much poorer than in the capital. One student said that in the middle of his senior year, they rounded up all the excellent, experienced teachers and sent them to work in a newly-opened school in AD, and that they were replaced with poor, inexperienced teachers. Those in the capital said that they didn't think there was much difference, based on what they'd heard from family members and friends who attended such schools.

When asked if they felt prepared for the CEPA exam, initially, they all responded positively. Then one of the students mentioned that they'd been given a 'CEPA vocabulary' list of 2000 words and were expected to know all the words by the end of their senior year. They talked about the frustration of studying that word list and then not recognizing more than 5-6 of those words in the CEPA exam itself! They said that the teachers stopped teaching the curriculum three months before the CEPA exam – one student said he'd even disposed of the English texts – and that three months was spent studying vocabulary and grammar anticipated to be included in the CEPA exam. There was a great deal of negativity about the whole experience and the poor way it seemed to be organised. One student pointed out that if the government wanted them to be prepared for such an exam, they should have started improving the English teaching situation and curriculum with primary students, and not obliged students who'd been exposed to a less-than-conscientious English education all their lives to sit for such an exam.

Students were asked if they felt that their CEPA scores correctly reflected their English ability 8 months ago (when they were finishing secondary school). Their response was overwhelmingly positive. They did add that it did not in any way reflect what they are now able to do in English – that they would do much better now if they could take it again.

The students were then asked if HCT was their first choice. For about half of the participants, it was. They cited the main reason for this is HCT's excellent reputation for quality delivery of courses, especially English. Others cited the great respect that a degree from HCT commands in the regional business community. "If you have a degree from HCT – any degree – any business will take you." And they also mentioned that HCT distinguishes itself because they have a job programme for student placement. "HCT helps us get good jobs, but if we went looking by ourselves with only a diploma, we wouldn't find a good job." They were very uncomplimentary about the national university. "Everyone with a degree from UAE University is sitting at home unemployed."

The ones who did not choose HCT first had either wanted a university degree (HCT provides only technical diplomas for students at the lowest intake level.), or had wanted to study abroad. One student seemed incensed about the requirements for this. He said, "I had wanted to go to Australia to study. I have family there. But they said that I had to pass TOEFL first. How the heck am I supposed to pass TOEFL when I struggled with CEPA?" He was bitter about the less-than-exemplary education, as he saw it, that he'd received at the preparatory and secondary levels.

The last topic covered in this group interview was about the motivation questions at the end of the questionnaire. The students were then asked if one of the reasons why they wanted to learn English was to understand English literature and art. Most felt that this was an interesting, but not absolutely necessary, benefit of learning English. They seemed to understand that learning a bit about the culture goes hand-in-hand with learning the language. They said, "It is good/interesting to learn things about the Western culture, but that is their culture (i.e. the culture which belongs to westerners). We have a culture of our own."

The last question produced some interesting responses: "Studying English is important to me because other people will respect me more if I have a knowledge of a foreign language."

They were not quite sure how to react to this question initially. Some hesitatingly agreed, citing the pride of family and the added respect one would expect at work if he spoke English well. Then one of the more vocal students interrupted and declared that respect is not dependent upon knowing another language. That even people with no second language and no high position command respect by the kind of person he or she is, by how that person treats others. He went on to say that his mother can neither read nor write, but that he respects her above anyone he knows. An explosion of agreement followed.

4.4.2 The Second Discussion: May, 2008

Length of Discussion: 60 minutes; Number of students attending: 15 (Three of those students originally contacted, but who did not attend the first interview, decided to be part of the second one.) The general subjects of discussion for the first group discussion, which was mostly about their experiences and opinions of the CEPA, were recalled as well as their secondary school experiences. The researcher explained that this discussion would concentrate on their experience here at college. This was important for the investigation of factors which appear to correlate positively with success.

The participants were asked about their feelings regarding the English instruction that they received at the college. One student said that there was much improvement. Another added that during this first year at college, they had learned more English than in the previous 12 years of schooling. Another student complained that government schools were well-known to provide poor language instruction. They were asked if any of them had graduated from a private school. One of them indicated that he had – an Arabic medium private school. It was mentioned that it has been estimated that 50% of Emiratis place their children in private schools rather than government schools. (This was presented as a fact by a secondary

school teacher here in Abu Dhabi during a morning of workshops organized by TESOL Arabia.) The researcher asked why. Short answer: English instruction.
Other responses included:
"It is known that marks can be manipulated in private schools."
"The fact that you pay money means that passing is virtually guaranteed."
"Still, everyone [whether in public or private schools] must take the government exams at the end of secondary. So just passing has become irrelevant."
"There are well-known and established private schools that offer a really good education."

They were asked to consider what would be their choice for their own children – in the future – private or public education. The overwhelming majority quickly replied "private". One student said that he would start his children in private school, and then transfer them to public at the preparatory stage. Others objected to his plan. Another student said that public schools are improving a lot. He mentioned the arrival of Western consultants and the revision of curricula.

Students were asked, "How do you feel about the quality of instruction of the other subjects you have in Diploma Foundations?" (Computer & Maths). Most said that Maths and computer were easy courses, but English was the toughest. One student said that Maths instruction in private schools was generally weak.

They were then asked if they were satisfied with the academic support that they are entitled to at the college outside of their actual class time. They were encouraged to voice their opinion about the academic guidance and support provided by the college (i.e.: Choosing a major for the upcoming year). They seemed to be generally satisfied with the support given them. One student said that he would have preferred a wider choice of majors. Three students said that they had wanted to choose Media/Communications, but that this choice is not available to Diploma Foundations students. Only two of the group discussion participants had taken advantage of the counseling available to them with different results. One was very positive, but the other was mostly negative, owing most likely to the fact that he'd been referred to the counselor for disciplinary action (excessive absenteeism).

The discussion got off-topic because one student expressed his opinion that the English ability of students in Higher Diploma was no better than Diploma Foundations. Others concurred. The researcher queried them about how this might be possible. One student gave two reasons: 1: Knowing someone in a position of power who could arrange this; and 2: Cheating on the CEPA.

Their opinion of the remedial assistance that the college makes available to them was mixed. One student said that the remedial assistance was excellent. When he was asked whether he meant from the teacher or from the special remediation centre in the college library, the students expressed real dissatisfaction with the centre. One student said that all they did was to load him down with photocopied exercises that he had to complete on his own. Several students complained that the English ability of the Maths tutor was so bad that they had been obliged to 'teach the teacher'.

Changing the subject, the students were asked how they felt about the assessment situation here at the college … the way it is organized and weighted? Their response was very positive and they expressed general satisfaction. They were asked if they'd felt well-prepared for the exams that they'd sat for. Their response was even more positive. One student went on to explain that everything was better: the way he'd been taught, the way he'd been prepared by the instructors for the exams, and was even happy with the exams themselves. "Better than what?" And many students responded with: "Better than the assessment situation in school." Why was the situation better here? Many responded strongly that the college's system was far better than their experience in school. They said that in college, even if they'd messed up in one exam, there was a chance to make up for a bad test mark with other assessments during the year. – If they don't do well the first time (in college), they have the chance to make it up in the tests that follow. They were asked to clarify what was different. They said that in school, they sat for monthly tests, and a mid-term and a final exam. Their first term marks counted for half of their final mark. They were not very clear about exactly how the high school system differs from the college's system, but most of them said that the college's system was much better. One student said: "In college, if you have studied for the exam, you will pass", apparently indicating that the same could not be said about secondary school. The students all responded affirmatively as to whether they were aware of the weighting scheme of the formal, summative assessments at the college.

"So far, you've managed to make it thus far to the final exams. What has helped you to make it to this point?" One responded, "Personal effort. Study." One student said that some students just relied on luck; that they do not study. Some other students, they said, had families who obliged them to study. One student said that having a positively supportive family was really important. Another said that some students were less motivated because they'd already found employment. He went on to say that it depends on what a person can be satisfied with. He said, "Some people can be content with 10,000 dhs/ month but others want much more. Those who want much more must have a degree."

At the end of the discussion, students were asked if there were any other comments they'd like to make. One student responded, "About the CEPA. We would prefer that only UAE nationals be allowed to take this exam." His opinion was that the CEPA was much easier when his brother took it before (when it was only for Emirati nationals). "They made the CEPA harder when they allowed non-national students to take the exam." Others said that perhaps it became more difficult so that all involved in secondary school English would take it more seriously. Other complaints followed about why the CEPA seemed more difficult, complaints which echoed complaints made in the first group discussion:

- NNS teachers taught them incorrect English
- Secondary school teaching methods are wrong.
- The English curriculum is not appropriate for the level required to do well in CEPA.
- The CEPA should not be an exit exam for secondary school.
- Students from areas outside the city feel disadvantaged by a poorer education.
- Private universities in the UAE still require applicants to sit for the TOEFL, and they do not recognize the CEPA as a valid assessment of English ability.

Some students felt that the assignment of students to DF and HD was not fair (also mentioned in comments in the questionnaire). Their responses to a query about their preparedness for the upcoming KCA exams were generally non-committal.

4.5 The Analysis of the Documentation

This section includes the results of statistical analyses performed on the data collected. It begins with the Pearson's correlations that were performed on quantitative data provided by the admissions information are explained. The section concludes with the third step of the statistical analyses, which was the Multiple Linear Regression.

4.5.1 Pearson's Correlations

Pearson's correlations were performed on the data collected for 0155 (Diploma Foundations, DF, and 070 (Higher Diploma Foundations, HD) for the academic year 07/08. These were completely separate comparisons: one for HD and one for DF. The data compared in this study were the following:

CEPA English Scores
College English Classwork Marks
College English Final Exam Marks
College English Course Marks
HS English Final Marks (HS=secondary schl)
College Maths Classwork Marks
College Maths Final Exam Mark
College Maths Course Marks

The rationale for choosing this statistical analysis was that it is useful for comparing scores to point out strong and weak dimensional relationships between dual, numerical sets of marks. Also, Pearson's correlation is used when the distribution is normal. (If the distribution is not normal, a Spearman-Rho correlation is used.) CEPA English scores were included since this examination is the focus of the research study. High School English Final marks were included because this mark is a substantial part of the colleges' calculated admissions Placement Index (PI) for each student. GSC scores were not included for several reasons. First, even though the GSC is part of the PI, it is a very minor part (16%). Second, some of the students were Science stream and some were Arts stream. Since the GSC is a composite score of the final exam marks for all the secondary school subjects, and the breakdown of each student's scores for each individual subject was not part of the admissions data supplied to the researcher, this rendered the GSC an untenable inclusion. All the final college English marks were included since a major focus of this study was to explore the predictive validity of the CEPA. Another purpose of the study was to explore what possible factors account for the CEPA's reported highly positive correlations. It was thought that an exploration of the three separate parts of students' final English marks might expose areas where the correlations were stronger or weaker. The students' three final Maths marks were included as a kind of control, the assumption being that the CEPA would highly correlate with English, but not with Maths. This was born out in the correlational analyses.

The results can be seen on the following pages in Tables 4.5 and 4.6, and in Diagrams 4.1 and 4.2. No high correlation was expected between the CEPA English scores and the college Maths scores, and none were found. Also as expected, there was a strong correlation between the CEPA English scores and students' final English course marks. The correlation was stronger for the HD cohort than the DF cohort, as expected as well. It would be difficult to determine within the parameters of this research study the actual reasons for this differ, but perhaps this is a result of the much narrower and higher score range of HD students' CEPA scores (160-180), whilst students with scores of anywhere from 159 downwards were accepted into Diploma Foundations. Scatterplot diagrams were initially performed in order to determine whether or not a bivariate analysis was justified. They provide a visual graphic of correlational strength: the closer together the dots are, and the more concentrated along a line, the stronger the correlation revealed.

KEY to abbreviations used in Diagrams 4.1 and 4.2, & Tables 4.5 and 4.6:
Mat_Final: students' final course marks for Maths
Mat_Fin_Exam: students' final exam results for Maths
Mat_CW: students' final course work results for Maths
EngFinal: students' final course marks for English
Fin_EngExam: students' final exam results for English
EngCW: students' final course work results for English
hs_eng: students' final course marks for High School English

Diagram 4.1: Scatterplot of Diploma Foundations (DF) Correlation Analysis Results

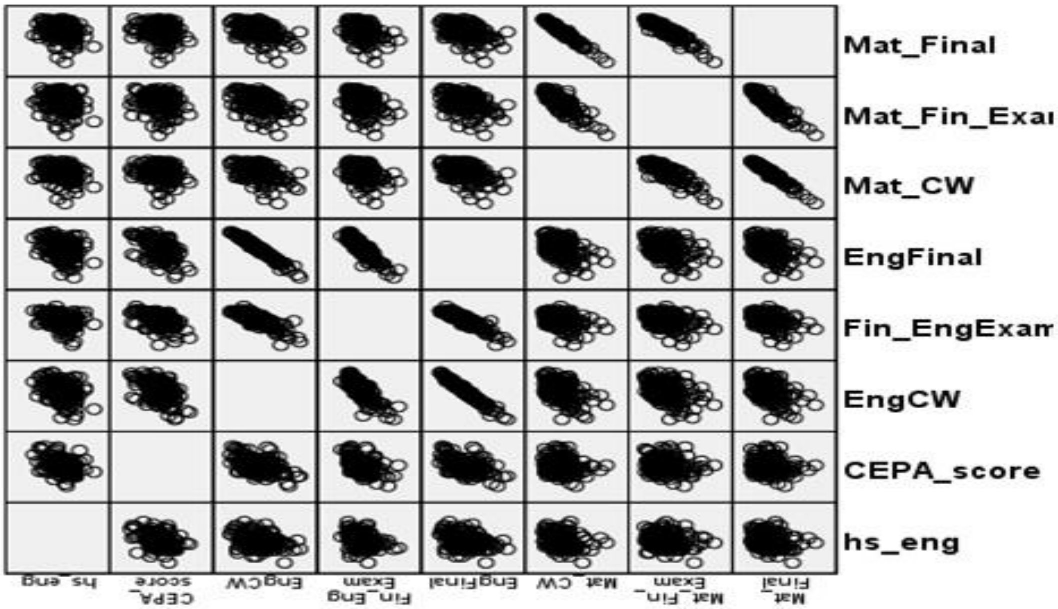


Table 4.5: Pearson's Correlation for Diploma Foundations (DF)

Diploma Foundations		Hs eng	CEPA score	Fin_Eng Exam	EngCW	Eng Final	Mat_Fin Exam	Mat CW	Mat Final
hs_eng	Pearson Corr	1	.386**	.296**	.370**	.357**	.169*	.247**	.229**
	Sig. (2-tailed)		.000	.000	.000	.000	.043	.003	.005
	N	146	146	145	146	146	144	146	146
CEPA score	Pearson Corr	**.386****	**1**	**.484****	**.468****	**.476****	**.140**	**.199***	**.194***
	Sig. (2-tailed)	**.000**		**.000**	**.000**	**.000**	**.092**	**.015**	**.018**
	N	**146**	**148**	**147**	**148**	**148**	**146**	**148**	**148**
Fin_Eng Exam	Pearson Corr	.296**	.484**	1	.833**	.913**	.339**	.377**	.381**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
	N	145	147	147	147	147	146	147	147
EngCW	Pearson Corr	.370**	.468**	.833**	1	.983**	.382**	.469**	.489**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
	N	146	148	147	148	148	146	148	148
EngFinal	Pearson Corr	.357**	.476**	.913**	.983**	1	.383**	.463**	.493**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
	N	146	148	147	148	148	146	148	148
Mat_Fin_ Exam	Pearson Corr	.169*	.140	.339**	.382**	.383**	1	.850**	.931**
	Sig. (2-tailed)	.043	.092	.000	.000	.000		.000	.000
	N	144	146	146	146	146	146	146	146
Mat_CW	Pearson Corr	.247**	.199*	.377**	.469**	.463**	.850**	1	.975**
	Sig. (2-tailed)	.003	.015	.000	.000	.000	.000		.000
	N	146	148	147	148	148	146	148	148
Mat_Final	Pearson Corr	.229**	.194*	.381**	.489**	.493**	.931**	.975**	1
	Sig. (2-tailed)	.005	.018	.000	.000	.000	.000	.000	
	N	146	148	147	148	148	146	148	148

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Diagram 4.2: Scatterplot of Higher Diploma Foundations (HD) Correlation Analysis Results

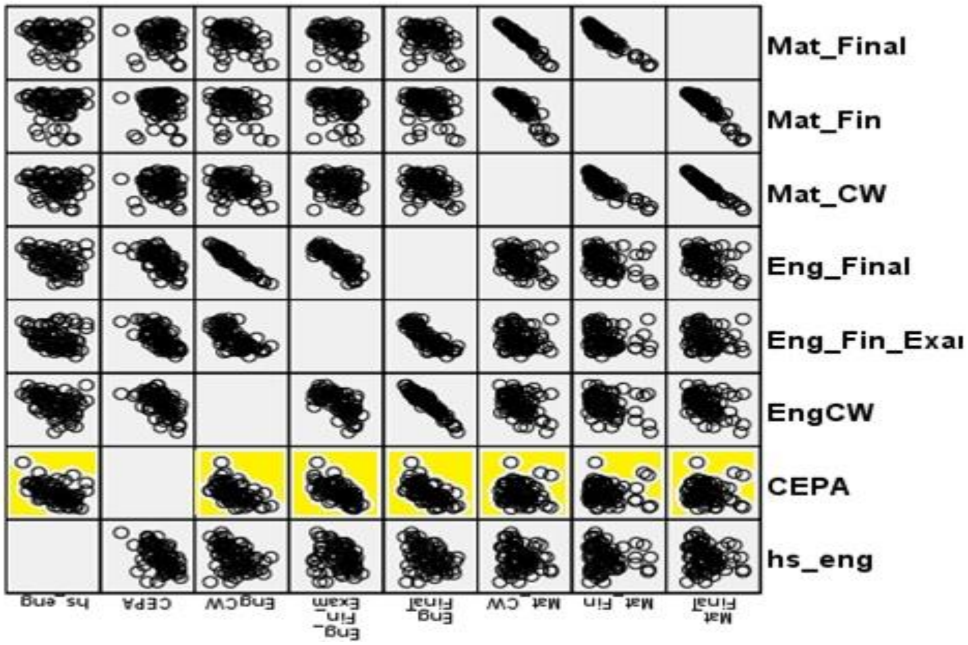


Table 4.6: Pearson's Correlation for Higher Diploma Foundations (HD)

Higher Diploma		hs_eng	CEPA	Eng CW	Eng_Fin Exam	Eng Final	Mat CW	Mat Fin	Mat Final
hs_eng	Pearson Corr	1	.515**	.391**	.238*	.370**	.168	.081	.138
	Sig. (2-tailed)		.000	.001	.041	.001	.151	.490	.240
	N	74	74	74	74	74	74	74	74
CEPA score	Pearson Corr	.515**	1	.564**	.679**	.652**	.096	.065	.086
	Sig. (2-tailed)	.000		.000	.000	.000	.418	.585	.468
	N	74	74	74	74	74	74	74	74
EngCW	Pearson Corr	.391**	.564**	1	.661**	.964**	.394**	.331**	.378**
	Sig. (2-tailed)	.001	.000		.000	.000	.001	.004	.001
	N	74	74	74	74	74	74	74	74
Eng_Fin_Exam	Pearson Corr	.238*	.679**	.661**	1	.837**	.190	.182	.191
	Sig. (2-tailed)	.041	.000	.000		.000	.106	.121	.104
	N	74	74	74	74	74	74	74	74
Eng_Final	Pearson Corr	.370**	.652**	.964**	.837**	1	.355**	.306**	.344**
	Sig. (2-tailed)	.001	.000	.000	.000		.002	.008	.003
	N	74	74	74	74	74	74	74	74
Mat_CW	Pearson Corr	.168	.096	.394**	.190	.355**	1	.911**	.987**
	Sig. (2-tailed)	.151	.418	.001	.106	.002		.000	.000
	N	74	74	74	74	74	74	74	74
Mat_Fin	Pearson Corr	.081	.065	.331**	.182	.306**	.911**	1	.965**
	Sig. (2-tailed)	.490	.585	.004	.121	.008	.000		.000
	N	74	74	74	74	74	74	74	74
Mat_Final	Pearson Corr	.138	.086	.378**	.191	.344**	.987**	.965**	1
	Sig. (2-tailed)	.240	.468	.001	.104	.003	.000	.000	
	N	74	74	74	74	74	74	74	74

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Tests of correlation are sometimes misused as tools of prediction. Establishing that correlation exists does not and should not imply a causal relationship. There may indeed exist a causal relationship, but this cannot be sufficiently determined by correlation alone. This is a limitation of the Pearson's Correlation Coefficient. It presents the relationship between two sets of variables, but it cannot predict. It also cannot compare multiple variables, nor can it be used with non-numerical data. This is why it was necessary to perform another statistical test: the Multiple Linear Regression (MLR).

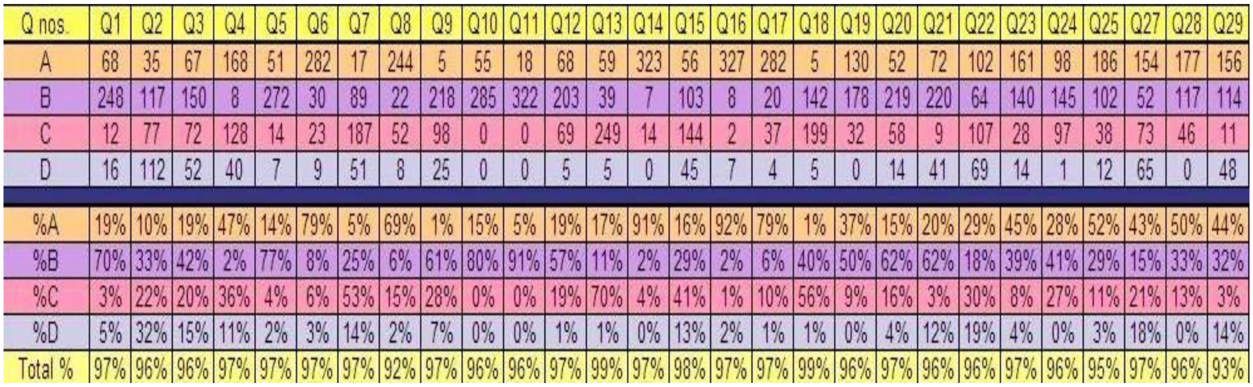
4.5.2 Multiple Linear Regression (MLR)

Before proceeding with a Multiple Linear Regression analysis, it is important to ascertain that the dependent variable for each test is normally distributed. The Kolgorov-Smirnov test was run on the final results for English and Maths and was discovered to be normal.

The MLR was chosen to statistically include and analyse the responses to the student questionnaire, their CEPA results and their final English and Maths marks at the end of their foundation year. The rationale for choosing the MLR is that it allows us to compare quantitative and qualitative data and to make predictions. It also enables the comparison of a multitude of variables. One of the goals of the research project was to explore a range of factors, identified by students themselves, the researcher and others who live in the UAE and work at the college to discover if any of these might also contribute to students' academic success as measured by their first year final marks. For this reason, the dependent variables were their final marks: English and Maths. The independent variables analysed were quantified responses from the students' questionnaires, their overall high school final marks and their CEPA scores. Variables excluded from the questionnaire data collected were either those that were irrelevant to this particular analysis, or those which were related to the first year final college marks themselves. It was not possible to include the latter, since the MLR is not designed to compare a variable to itself.

Diagram 4.3 shows the scanned results of the student questionnaire. The responses do not add up to 100% because the students did not answer all of the questions, which was unfortunate, but ultimately their right. This was also the reason for the post-questionnaire truncation of the student cohort. Of the approximately 350 completed questionnaires, which comprised the entire cohort, 150 DF students and 75 HD students chose to identify themselves. These students' responses were the only ones included in the MLR since it was only possible to compare questionnaire responses and CEPA results with final results for students who named themselves. Still, their names were only used to complete the statistical comparison, and then the records were destroyed.

Diagram 4.3: Excel representation of the scanned student questionnaire results



Q nos.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q27	Q28	Q29
A	68	35	67	168	51	282	17	244	5	55	18	68	59	323	56	327	282	5	130	52	72	102	161	98	186	154	177	156
B	248	117	150	8	272	30	89	22	218	285	322	203	39	7	103	8	20	142	178	219	220	64	140	145	102	52	117	114
C	12	77	72	128	14	23	187	52	98	0	0	69	249	14	144	2	37	199	32	58	9	107	28	97	38	73	46	11
D	16	112	52	40	7	9	51	8	25	0	0	5	5	0	45	7	4	5	0	14	41	69	14	1	12	65	0	48
%A	19%	10%	19%	47%	14%	79%	5%	69%	1%	15%	5%	19%	17%	91%	16%	92%	79%	1%	37%	15%	20%	29%	45%	28%	52%	43%	50%	44%
%B	70%	33%	42%	2%	77%	8%	25%	6%	61%	80%	91%	57%	11%	2%	29%	2%	6%	40%	50%	62%	62%	18%	39%	41%	29%	15%	33%	32%
%C	3%	22%	20%	36%	4%	6%	53%	15%	28%	0%	0%	19%	70%	4%	41%	1%	10%	56%	9%	16%	3%	30%	8%	27%	11%	21%	13%	3%
%D	5%	32%	15%	11%	2%	3%	14%	2%	7%	0%	0%	1%	1%	0%	13%	2%	1%	1%	0%	4%	12%	19%	4%	0%	3%	18%	0%	14%
Total %	97%	96%	96%	97%	97%	97%	97%	92%	97%	96%	96%	97%	99%	97%	98%	97%	97%	99%	96%	97%	96%	96%	97%	96%	95%	97%	96%	93%

Only data from AY 07/08 were included because that was the year the student questionnaire was administered, and the year in which admission was granted to virtually all who wished to join. The stepwise regression analysis method was chosen because this analyses the effect of variables upon each other. Within the stepwise method, it is possible to produce a list of factors which it has found to be statistically significant. A removal criterion automatically removes any predictor that does not prove to be statistically significant, and then the model is re-estimated for all the remaining predictors. For example, this analysis revealed that the amount of time they took to get to college in the morning was significant. This list of factors can then be used to develop a model to test the reliability of the MLR results. The basic MLR model is: “$Y_1 = (b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n) + \mathcal{E}_i$. Y is the outcome variable [- the constant], b1 is the coefficient of the first predictor (X1), b2 is the coefficient of the second predictor (X2), bn is the coefficient of the nth predictor (Xn), and is the difference between the predicted and the observed value of Y for the nth participant” (Field, 2005: 157).

The variables included in the MLR were those which related most closely to the issue to be explored: whether there were identifiable factors that might have potential as indicators of academic success. Several variables were not included because they were part of a composite mark. For example, the final course marks – whether English or Maths – are a composite of course work marks, and final exam marks. On the other hand, composite high school marks (the GSC) was not included since it is an overall average of several course subjects for which there is no similar course in the college (i.e. Arabic). Also, variables supplied by the Ministry of Education were deemed more reliable than students' recollections and more precise (i.e. the CEPA marks). For this reason, data like this in the admissions file was entered and not from the questionnaire. Had these admissions files been available to the researcher before the questionnaire was administered, these redundant items could have been removed from the questionnaire before administration to the students.

The stepwise method has been criticised for several reasons, but the one most relevant to be cautious of in this particular study is that the regression results will often fit much better in sample than they do on new out-of-sample data. A way to test for errors in models created by the stepwise regression model is to not rely on the model's F-statistic, significance, or multiple-r, but instead assess the model against a set of data that was not used to create the model (Mark and Goldberg, 2001: 92). This can be effectively done by using as a model the identified predictors based on most of the dataset available (i.e. 90%). Inserted on the next page is part of the output of the MLR on 90% of the DF data.

The model was then applied to the remaining 10% of the dataset to assess the accuracy of the model. The results of the analyses revealed the relative strength of the predictive validity of various contributory factors for this cohort, based on the factors chosen by the MLR as predictors. The results of the analyses of the MLR predictions for this research are shown in Tables 4.7 and 4.8. A Pearson's correlation was performed on the 10% of the total sample that left out of the main MLR. This was a straight comparison between the students' actual final marks and what the MLR model predicted their marks would be, given the variables for each student. (The results of this last analysis may be found on the following pages.)

Insert 4.1: 0155 (DF): 90% MLR stepwise results – English Final Dependent Variable:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14.743	12.712		-1.160	.248
	CEPA_score	.558	.084	.490	6.669	.000
2	(Constant)	-17.331	12.248		-1.415	.159
	CEPA_score	.559	.080	.491	6.954	.000
	Q38dummy1	5.018	1.423	.249	3.527	.001
3	(Constant)	-19.944	11.855		-1.682	.095
	CEPA_score	.570	.078	.500	7.332	.000
	Q38dummy1	5.446	1.380	.270	3.946	.000
	Q35dummy3	7.491	2.249	.228	3.330	.001
4	(Constant)	-22.193	11.621		-1.910	.058
	CEPA_score	.589	.076	.518	7.722	.000
	Q38dummy1	5.089	1.356	.253	3.754	.000
	Q35dummy3	8.484	2.230	.259	3.805	.000
	Q6dummy2	-5.922	2.178	-.186	-2.719	.007
5	(Constant)	-22.774	11.354		-2.006	.047
	CEPA_score	.601	.075	.528	8.046	.000
	Q38dummy1	4.869	1.327	.242	3.669	.000
	Q35dummy3	8.549	2.178	.261	3.925	.000
	Q6dummy2	-6.007	2.127	-.188	-2.823	.005
	Q27dummy3	-4.178	1.514	-.180	-2.759	.007
6	(Constant)	-30.659	11.664		-2.629	.010
	CEPA_score	.525	.080	.461	6.535	.000
	Q38dummy1	4.749	1.306	.236	3.635	.000
	Q35dummy3	8.570	2.143	.261	3.999	.000
	Q6dummy2	-5.939	2.093	-.186	-2.837	.005
	Q27dummy3	-3.590	1.511	-.155	-2.376	.019
	hs_eng	.287	.122	.166	2.350	.020
7	(Constant)	-33.844	11.591		-2.920	.004
	CEPA_score	.541	.080	.475	6.807	.000
	Q38dummy1	5.247	1.308	.260	4.012	.000
	Q35dummy3	8.892	2.118	.271	4.198	.000
	Q6dummy2	-5.670	2.068	-.178	-2.742	.007
	Q27dummy3	-3.570	1.490	-.154	-2.397	.018
	hs_eng	.289	.120	.167	2.399	.018
	Q9dummy1	8.723	3.946	.143	2.211	.029
8	(Constant)	-34.402	11.437		-3.008	.003
	CEPA_score	.547	.079	.480	6.966	.000
	Q38dummy1	4.968	1.296	.247	3.832	.000
	Q35dummy3	8.801	2.090	.268	4.211	.000
	Q6dummy2	-6.125	2.051	-.192	-2.987	.003
	Q27dummy3	-3.814	1.474	-.165	-2.588	.011
	hs_eng	.281	.119	.163	2.367	.019
	Q9dummy1	9.029	3.895	.148	2.318	.022
	Q15dummy1	4.504	2.074	.137	2.171	.032

a. Dependent Variable: EngFinal

Model Summary for 0155 90% - Eng Final Dependent Variable

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.490[a]	.240	.234	8.82803
2	.549[b]	.302	.292	8.49034
3	.594[c]	.353	.339	8.20000
4	.622[d]	.386	.369	8.01766
5	.647[e]	.419	.397	7.83218
6	.664[f]	.441	.417	7.70601
7	.679[g]	.461	.433	7.59819
8	.692[h]	.479	.448	7.49576

a. Predictors: (Constant), CEPA_score
b. Predictors: (Constant), CEPA_score, Q38dummy1
c. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3
d. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3, Q6dummy2
e. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3, Q6dummy2, Q27dummy3
f. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3, Q6dummy2, Q27dummy3, hs_eng
g. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3, Q6dummy2, Q27dummy3, hs_eng, Q9dummy1
h. Predictors: (Constant), CEPA_score, Q38dummy1, Q35dummy3, Q6dummy2, Q27dummy3, hs_eng, Q9dummy1, Q15dum

The predictors identified by the MLR for the Diploma Foundations (DF) students' first year final marks common to both English and Maths were the CEPA scores, and their respective high school final marks. There were 13 questionnaire item responses that the MLR discovered were statistically significant with regards to DF students' first college year final marks. (See appendices B and H.) Common to both analyses were items 6a and b (mother's nationality), 27c (preferred choice of college major) and 35d (a lack of interest in learning English for art and literature). Statistically significant item responses for the final DF English marks were 9a (future employment preference), 15a (taking less than 15 minutes to get to college) and 38a (a desire to study English to gain more respect).

Table 4.7: DF results of the MLR analysis of factors with 10% of sample

10% Analysed with Model produced by MLR					
#	Student's Research ID #	**Actual English Final Mark**	**MLR Predicted English Mark**	**Actual Maths Final Mark**	**MLR Predicted Maths Mark**
1	1001	81.38	79.31	88.83	87.53
2	1003	73.49	79.54	73.75	76.97
3	1004	78.84	77.65	65.84	65.31
4	1006	87.04	82.19	90.63	89.98
5	1007	80.16	81.83	92.60	92.60
6	1010	83.19	80.39	91.94	91.16
7	1015	83.31	82.96	77.44	76.33
8	1016	81.28	83.41	69.85	69.85
9	1018	83.72	85.78	88.16	89.45
10	1019	77.34	77.34	68.52	68.92
11	1040	75.86	75.71	78.73	77.34
12	1047	71.88	72.92	80.05	80.79
13	1048	70.73	74.71	73.61	72.31
14	1049	79.94	75.31	79.42	80.64
15	1053	86.14	85.19	80.07	80.18

➔ ***Pearson's Correlation for DF English mark: r=0.798***
➔ ***Pearson's Correlation for DF Maths mark: r=0.99***

Table 4.8: HD results of the MLR analysis of factors with 10% of sample

10% Analysed with Model produced by MLR					
#	Student's Research ID #	Actual English Final Mark	MLR Predicted English Mark	Actual Maths Final Mark	MLR Predicted Maths Mark
1	2003	78	76.81	89	86.52
2	2005	71	76.03	77	79.53
3	2006	80	78.47	82	82.23
4	2007	82	79.16	80	79.10
5	2008	76	77.25	81	80.74
6	2010	77	79.87	88	88.36
7	2011	77	75.60	91	91.82
8	2012	79	76.18	76	76.05

➔ ***Pearson's Correlation for HD English mark: r=0.469***

➔ ***Pearson's Correlation for HD Maths mark: r=0.968***

Statistically significant item responses for the final DF Maths marks were 4a (father's state of employment), 7c (father's primary job), 8c (mother's primary job), 12c (more than 5 siblings), 17c (type of high school attended), 19b (followed Arts stream in high school) and 25c (the feeling that he was not placed correctly in Foundations). The predictors identified by the MLR for the Higher Diploma Foundations (HD) students' first year final marks common to both English and Maths were their CEPA scores. High school final Maths marks were also found to be statistically significant for HD students' first year Maths final marks. There were five questionnaire item responses that the MLR discovered were statistically significant with regards to DF students' first college year final marks. (See appendices B and H.) Common to both the English and Maths final marks analyses were items 17 (b and c: type of high school attended), and 19 (b and c: stream he followed in high school) respectively. The only other statistically significant item response for the final HD English marks was 3a (the illiteracy of his mother). Statistically significant item responses for the final HD Maths marks were 7b (father's primary job) and 9c (student's future employment preference). These results will be further discussed in the following chapter.

4.6 Summary

This chapter presented and explained the data collected for this study. The focus of this chapter was to provide a thorough overview of the data collected. The discussion and analysis of this data was reserved for the following chapter. After describing the development, piloting, revision and administration of the student questionnaire, the responses themselves were presented. These responses revealed a largely homogenous group of students with regards to their backgrounds, educational experiences, and aspirations. For example, apart from the fact that they were all young Emirati men, the majority of them have parents with little or no education. Most of them come from large families and continue to live with their families whilst in college. Additionally, most of these first year college students are graduates of public secondary schools, which often have severe problems in classroom management, control of cheating and quality of instruction. (It was largely because of these educational deficiencies that the initiative to develop the CEPA was encouraged and supported by the government.) Additionally, the results of statistical analyses of the predictive validity of the CEPA and the predictive validity of the factors identified by students and others as either positively or negatively affecting their ability to succeed academically were presented.

In the following chapter, the data presented here is discussed and possible explanations examined. An exploratory intent motivated the data collection process: the desire to explore the predictive validity of an examination on a homogenous group of students who were all given the opportunity to study at this college, and the desire to explore how valid were the factors the students identified as affecting their chances of academic success, as well as others intimately involved with the college specifically and the target culture in general.

5: Discussion and Analysis

5.1 Introduction

The aim of this chapter is to look at the research project as a whole in order to reflect on the findings of the case study and what has been learned. It is hoped that this will contribute to our understanding of language proficiency test score use in the context of college admissions and placement decisions. It also presents an opportunity to reflect upon the research methodology used and the approach taken for the data analysis in order to discuss outcomes of the research project and process.

5.2 Overview of the Study: The Literature

The literature review covered several elements important to the background of the research. First, whilst not actually a focus of the study, the dilemma of the confusing terminology of applied linguistics in general and testing issues specifically was briefly addressed. If one of the intents of action research based in participant data gathering (which often involving case studies) is to build up a body of information from which universal conclusions may be drawn, it would be useful and efficacious to have core terms which are universally understood and applied.

One of the terms of Applied Linguistics which has undergone perhaps the most discussion is test validity. The review of the literature on this topic began with early scholarly work on the different types of validity, beginning with the pioneer work of Cronbach and Meehl (1955), proceeding to the point where we are now, and that is a generally accepted view of test validity as comprehensive and unified, as described by Messick (1989, 1996), Alderson (1991), Bachman (2000), Fulcher and Davidson (2007) and others. Other developments in the discussion of validity were also reviewed, including the notion of Kane's Argument-Based Validation (2004) and Lissitz and Samuel's (2007) (widely criticised) proposal for a fundamental change in not only our understanding of unified validity, but also the terminology which describes different facets of validity itself in an attempt to counter the most problematic part of Messick's Unified Validity theory: that of verifying the construct. Although the debate about how to establish validity and what it means in practical terms vis a vis test development and validation will most likely continue, there does seem to be general acceptance that not only are all the types of validity – construct, content, criterion, predictive – intertwined and necessary (to varying degrees), but that even reliability has been suggested as being a part of a test's validity as an instrument (Alderson, 1991). Because it is a central focus of the present study, the literature about predictive validity was reviewed in detail. Criticisms of the notion were noted, as well as acknowledging its popularity and its pervasiveness. Wrapping up the section on test terminology, literature about the advantages and disadvantages of norm and criterion referencing was reviewed.

The item types and scoring systems that are included in the CEPA were then discussed and reviewed. The literature about multiple choice questions and its advantages and disadvantages revealed strong criticism of its apparent 'objectivity' and the difficulty of developing well-written items. This section also covered explanations of and discussions about scaling scores for tests, and band scoring of writing tests. This included an overview of intra- and inter-rater reliability. The review of the literature about band scoring included Lumley's (2006) comprehensive model of the rating process.

More subjective, but nevertheless major considerations in test design, development and administration were then reviewed. This included information and opinions about issues of practicality in test development and administration, testing ethics and the desire for fairness, and the affective domain of the test takers themselves.

The review of the literature concluded with a synopsis of predictive validity studies comparable to this one (which is about the CEPA) in terms of focus and scope, and studies about different experiences around the world in regionally-developed English proficiency examinations.

This review of the literature on the topics associated with testing in general and with issues specifically related to the CEPA was necessarily wide in scope, and attempted to be reflective of the role tests play in our societies in general. This review laid the foundation for the exploration of the core issues of the current study: reliability and validity (and more specifically, predictive validity) of the CEPA, the uses of CEPA, the characteristics of the CEPA test-takers and how the experience of the development of the CEPA might benefit the wider testing community with regards to the production of regionally- developed proficiency exams.

5.3 Overview of the Study: The Research Design

The paradigmatic focus of this study is a mixed method (MM) one, combining qualitative and quantitative data gathering techniques in the investigation of a case study. Tashakkori and Teddlie (2003: 713, as noted in Chapter 3) described the philosophical orientation of mixed method studies as a pragmatic one. They defined pragmatism as "a deconstructive paradigm that debunks concepts such as 'truth' and 'reality' and focuses instead on 'what works' as the truth regarding the research questions under investigation" (ibid.). Figure 3.1 (p. 90) outlines the conceptual framework for the study. The framework was labeled a "convergent sequential design" based on the descriptions of various research study designs described by Cresswell and Plano Clark (2011) and Tashakkori and Teddlie (2009). As explained, the study is sequential in that the steps of the data gathering stage flowed sequentially from one to the other. It is described as convergent because the results of the qualitative data gathering and the quantitative data gathering were merged in the final analysis stage of the research design.

The sample investigated was the entire cohort (n=347) of first-year foundations students at a vocational college in the United Arab Emirates. This particular cohort was unusual in that a special set of circumstances enabled the college to offer admission that academic year (AY 2007-8) to all who were interested in joining; meaning an almost unheard of non-truncated sample. After the administration of the student questionnaire to these 347 students, the first step of the data gathering process, it became necessary to remove from the analysis some of the respondents, as previously mentioned and explained in Chapter 4.

Analytical procedures employed in this research study included the scanning and tabulation of the student questionnaire responses, the translation and organisation by topic of the free response sections of the questionnaire, and the transcription of all the interviews with college staff and administration officials involved in the assessment of students, and/or the admissions process. Data from the students' admissions files, quantified questionnaire responses and their final first year college English and Maths marks were analysed using different statistical methods. Two-tailed Pearson's correlations were performed to determine the strength of the correlation between the students' CEPA scores and their final first year college English and Maths marks. Then a Kolmogorov-Smirnov test was performed on the dependent variables from the two levels of the 2007/8 student cohort (DF and HD) as a preliminary step to confirm the normality of distribution of the data before doing the final analysis. This was a regression analysis, which was performed to identify predictor factors which might be used as potential performance indicators.

5.4 Research Questions and Findings

The focus of this investigation, as articulated in the thesis question is:

Is strong predictive validity possible for regionally-developed English proficiency examinations used for admissions and/or placement decisions?

In an attempt to answer the thesis question as thoroughly as possible, three research questions were generated. They are presented and discussed in the following sections.

5.4.1 Research Question (A): Is this locally-produced English proficiency examination a useful statement of competence and a predictor of academic achievement?

The three sub-questions for this first research question take into consideration several topics crucial to unpacking and answering this research question. The first is about determining the validity and reliability of the CEPA. The second involves an issue related to validity and reliability, which is test use. It asks how CEPA is used in making admissions and placement decisions. Finally, the last sub-question considers issues regarding the determination of the CEPA's predictive validity.

Sub-question One: Is the CEPA a valid estimate of students' English ability before admission?

On the face of it, this may seem to some actually two questions, but if we agree that a test which is not reliable cannot be valid and vice versa, we acknowledge that validity and reliability are essentially two sides of the same coin and both are a way of answering "How 'good' is it?" This issue was considered at length in Chapter 2 (for example, see Alderson's discussion of this on page 25 in which he stated that it doesn't really matter what we call these elements of reliability and validity as long as they can be demonstrated to exist). As with many other assessment tools, the response to "how good is it?" differs depending on the individuals or groups responding.

What the Foundations supervisors and co-ordinators said:

The Higher Diploma supervisor was critical of the band marks of the writing section. She mentioned requesting the papers of a couple of borderline students to mark them herself. She said that she would have awarded the papers at least a band mark more than they'd received. Because of this, she also felt that the CEPA score would be more reflective of true English ability if the essay scores were not included. For his part, the Higher Diploma Foundations co-ordinator simply stated that they never pay attention to whether CEPA marks correctly match teachers' perceptions of students' abilities or not. "Once we get a class list, that's it." The Diploma Foundations supervisor had nothing specific to say about this, but the DF co-ordinator suspected that the CEPA results were inflated due to coaching. This apparent 'inflation' had led to the discovery of errors in class section streaming for DF, and difficulties in the preparation process for the course (i.e. in tailoring learning objectives to students' abilities).

What the central administrator said:

The central administrator said that a CEPA score of 165 and above would be "a good indicator of future success of the HD Foundations [final exams]". And even though he said that the CEPA does not "provide us with a measure of skills", he spoke of the value of the CEPA as "an admissions tool" and as "a score for admissions purposes". Specifically, he said that it's "useful in terms of admissions and students' future progress".

What the CEPA supervisor said:

The CEPA supervisor said that they'd achieved construct validity by comparing the different item types included with "theories of language". He gave an example of using Bachman's framework to "see grammatical competence and vocabulary" (see p. 152). He also mentioned that the CEPA possesses concurrent validity because they'd done studies wherein students who took the CEPA also took the TOEFL and IELTS. He said, "We would expect high ability students in one test to be high ability students in the other test and vice versa. And that's the case and we do have strong correlations with both the TOEFL and IELTS". He also stated that the CEPA is a good predictor of language ability. As for the reliability of CEPA, the supervisor couched his assertions in unreal conditionals: "If it wasn't reliable, then you wouldn't see strong correlations in the measures of language ability – but we do. If it wasn't reliable, you would see low internal consistency in the test itself. We have very high internal consistency." He also explained the internal test statistics they'd collected for the CEPA: "We have the item-level information on the test data… like the KR20 or the Cronbach-Alpha which measures the extent to which you have inter-item correlations. Typically in high-stakes tests, you want something that is over .90. We're consistently at .95 or .96 in all of the tests. We've *never* been below .94. And that's as good as anything the

TOEFL ever gets." Additionally, Brown and Jaquith (in O'Sullivan, 2011: 260) declared that since 2003 "the CEPA has been committed to the appropriate use of available technology, [priding themselves] on being at the cutting edge of modern language testing".

Many of the comments about the validity and reliability of the CEPA seem somewhat imprecise, and there are several areas/issues that they refer to or speak about which might rightly be criticised for being weak or uninformed, but what the testers do all share is an intuition about the effectiveness of the assessments and some general, intuitive idea about the students they are assessing and/or teaching. This intuition demonstrates, as Messick (1998: 37), Broadfoot (1984:123) and others (Brown, 1990: 42, McNamara, 2001: 337) have noted about the development of tests, that test development is actually more akin to a skill or even an art, than it is a 'science'. Tests are developed by people – people who, consciously or unconsciously, bring to the items they develop certain socio-cultural backgrounds and experiences. Tests cannot be entirely divorced from this human element, so some have argued that even the idea of test validity is relative – relative to the educational situation, the stakeholders, the political-economic climate, etc. What experienced test developers share is an appreciation of, and an intuition for, those for whom tests are being developed. Logically, it would seem that this should be even truer for locally- or regionally-developed examinations, but there is little empirical evidence.

Is the CEPA a valid and reliable estimate of students' English ability at the end of their secondary education? One might venture a qualified and cautious "Yes". All indications are that it is, based on the empirical data collected by NAPO (National Assessment and Placement Office, the organisation responsible for producing the CEPA, among other things), as explained by the CEPA supervisor and the assertions of the college's central administrator in the semi-structured interviews, and the statistical data generated by this research study.

Sub-question two: "How are the results of the CEPA exam used in college admissions and placement decisions?"

The CEPA exam was originally designed as a post-admissions placement tool, developed jointly by the two UAE universities and HCT. This was its original purpose, starting in 2002. It was a low-stakes placement assessment. In 2006, the decision was taken (unclear by whom) to not only continue to use the CEPA results to stream admitted students, but also as a major part of university admissions criteria. As explained by Brown and Jaquith (in O'Sullivan, 2011: 253), this necessitated several changes to the CEPA, not least of which was the training and remuneration of raters (which had not previously been done) and the development of online marking to expedite the marking of the writing section. In 2007, it was decided to make CEPA an exit requirement for public high schools in the UAE in addition to its other uses.

In the interview with the college registrar for this study, he touched on the admissions process employed by all 3 of the national tertiary institutions (this college, and 2 universities), as mandated by the Ministry of Education. The UAE has many colleges and universities, but only 3 of them are specifically for UAE nationals. As previously mentioned, the study was conducted on one of the 14 campuses of a UAE vocational college. Each major metropolitan area has separate men's and women's campuses for this college. As discovered in the interview discussion, because essentially the same process is employed by all three institutions, there was no longer any need for the research sub-question which had sought to consider the different admissions procedures of the three national tertiary institutions.

The college registrar revealed that admissions data about students is supplied to them from the Ministry of Education, and this is used to make a calculation. At the Higher Colleges of Technology, this is called the Placement Index (PI). As mentioned earlier, the PI is calculated by combining 30% of the GSC English score (high school), 16% of the GSC score (overall high school average) and 54% of the CEPA English score, in addition to a separate minimum required band score for the CEPA writing element. As the registrar explained, "The PI last year for Diploma [Foundations] was a minimum of 29 and above. For Higher Diploma, a student needs a minimum of 71. A minimum of 71 to be considered, you know. Suppose you got a PI of 71, but still your writing band is too low? It doesn't work. You need a writing band for Higher Diploma of 3.5 and above."

In addition to taking into consideration the academic standing of the applicants, college officials must also take into account other factors – like all colleges and universities do – which have nothing to do with the applicants themselves, but with the ability of the college to accommodate them. These factors include classroom space, the number of instructors and preferred teacher-student ratios. In the interview, the Registrar did an excellent job of detailing this part of the process at HCT (See pages 116-117.).

When the researcher asked the Registrar about the percentage of all the students accepted who apply to the college, at first he strongly denied that it was 100%, as reported by the DF co-ordinator, then he seemed to contradict himself by saying that the college does indeed accept more than just the target number if, for example (as happened in 2007), they received extra money for the college: "… At certain times, the college gives you extra budget [money]. Let's say for the second semester whoever you can take, you take as many of the waiting students as possible. Sometimes they are of a low level, so we don't take them to Higher Diploma. We take them into the Diploma Foundations. Diploma is kind of trying to make people literate, you know. Just to [establish] literacy. We're not going to take them into Higher Diploma or Bachelors, we're taking them to a Diploma. And if they improve their level, [there is no reason why they cannot advance to a higher level]. We should give a chance … the belief of the Colleges is [to provide opportunities.] ... So, that's what we do every year. We give a chance to those who did not have a chance before – the students of this academic year. And that's why they say we're taking 100%. We are not taking 100%. And not even 100% will show up because many of them are not interested, many of them do not understand the importance of the college". This stated aim of only establishing literacy in English may be a possible contributing factor to the lower correlation coefficient results of Diploma Foundations students.

A fundamental issue which was only been superficially alluded to by those interviewed for this study is the question of appropriate test use. Shohamy (2001: 162) emphasised this, saying, "Testers must realize that much of the strength of tests lies not only in the technical quality, but in their use in social and political dimensions. Studies of the use of tests, as part of text validation on an ongoing basis, are essential for the integrity of the profession". As stated by Bachman and Palmer (2005:17) at the beginning of a lengthy chapter on the topic, "The most important consideration in designing and developing a language test is the use for which it is intended". Bachman (2008: 53-79 and 280) defined the four major uses of language tests as being selection, placement, diagnosis and evaluation. By using the CEPA exam as a placement tool, admissions criteria and an exit requirement, it would seem that it has taken on the questionable role of all four uses. What is not at all clear is to what extent these additional uses of the CEPA results are justified or even valid.

"The purpose of a performance examination is to infer candidate abilities that go beyond the particular sample of tasks, items and judges encountered. Whether the goal is to make reproducible pass/fail decisions or to position candidates according to demonstrated ability, the performance examination must measure candidate ability consistently" (Lunz & Wright, 1997, in Weir, 2005: 27). Whatever the CEPA is measuring, it does seem to measure it consistently, as reported by the interviewed CEPA supervisor (See page 154.), but it remains unclear whether the consistency of results can justify all the uses to which it is being put. As Bailey (1998: 2), "The main purpose of language assessment is to help us gain the information we need about our students' abilities and to do so in a manner that is appropriate, consistent and conducive to learning". Questions about the CEPA's use appropriacy and its conduciveness to learning remain.

The research found that students' CEPA scores fulfill several functions. First, beginning in 2007, sitting for the CEPA became a requirement for completion of secondary education in the UAE. Second, since 2006 it has been used as part of the criteria for offering admission to prospective students. Students who score above 180 (and who have a high PI) will be admitted directly into degree programmes. Those who score less than 180 are admitted to the college, but must complete a year of 'foundation' courses to strengthen basic skills in English, Maths and computer. Third, the CEPA is used as part of the placement criteria into these foundations courses – either Higher Diploma Foundations or Diploma Foundations. The PI formula is the determining factor for admissions and placement, along with students' band score from the CEPA writing section at HCT. However, since the upper and lower ranges of the PI was not divulged to the researcher, it is not possible to determine, or to comment upon, the extent to which it defines what they have determined is a student's ability to succeed at HCT. To put it another way, the Registrar indicated that students with a PI of 29-70 are placed into Diploma Foundations. Is this out of 100 or 500? Obviously, it matters and is important to know.

One final point needs to be made: the intent for the development of the CEPA exam was never to replace the TOEFL or IELTS examinations. Ghazali's (2008) equivalency study (page 156), along with the CEPA supervisor's corroboration of the apparent correctness of those results (page 157) might possibly have revealed some basis for comparison of these different instruments, but this is actually not all that important, nor even relevant. The CEPA fulfills a different function than IELTS or TOEFL. It's not replacing these other tests, but apparently performing reasonably well for its original intended purpose: to assist in the admissions and placement process at government colleges and universities in the UAE. (Whether its use as an exit exam for high school is justified or not is not a focus of this study, but this should be investigated as well.)

Sub-question three: "What is the predictive validity of this English proficiency exam in gauging the ability of students to progress successfully in English courses and Maths courses in which the medium of instruction is English?"

The data seem to corroborate the assertions made by the individuals interviewed (as reported in the previous chapter and in the previous section) that the CEPA does successfully predict students' ability to complete their first college year. Two of the interviewees (the HD supervisor and the DF coordinator) voiced concern over the reliability of the CEPA scores themselves. The HD supervisor spoke about her skepticism with regards to the reliability of the band scores for the writing section of the CEPA. The DF coordinator expressed some frustration with inaccuracy due to intensive coaching for the CEPA exam. In addition, referring to what Prof. Charles Alderson whimsically referred to as a "virgin cohort" (personal conversation, July, 2005), the year in which the data was collected – 2007 – was the year, as indicated in the interviews, that almost no student was turned down for admission. HCT admitted almost 100% of recent secondary graduates who indicated their interest in enrolling with the college (the only constraint being physical space). And in any case, (as also revealed in the interviews) in general, it is not unusual for HCT to grant admission to over 90% of those who apply. This presented a unique opportunity to investigate predictive validity without the so-called fatal flaw of the truncated sample from which almost all other predictive validity studies suffer.

Results of Pearson's correlations

Before proceeding, it must be reiterated that it is not possible to rely on the results of Pearson's product correlations to predict future performance with any certainty, since a causal relationship cannot be established using this statistical method. Nevertheless, correlations can present corroboratory evidence, supporting the findings of other statistical analyses. The Pearson's correlations conducted on the data collected by the researcher revealed a strong, positive relationship between the results of CEPA and students' first year final marks in English: .476 for DF students (0155), and .652 for HD students (070) [statistical significance at 0.01 level]). The correlations between CEPA English scores and students' final first college year Maths marks, while statistically significant, were much weaker: .194 for DF students (0155), and .086 for HD students (070) [statistical significance at 0.01 level]), a result which was anticipated.

Even though the analyses revealed a strong correlation between CEPA and students' final first college year English marks, this was lower than what had previously been posted on the college's website in the Academic Services section in 2005 – correlations so high, in fact, that they served to encourage the researcher to initiate this study to investigate the phenomenon. In an email exchange with the Assessment Co-ordinator in February of 2008, he asserted different results (not referring at all to the 2004/5 results, which had by 2008 disappeared from the website). He said, "Look at the start of Item 3 in the KCA reports. I don't think it was ever that high:
2006, 070 [HD] and CEPA was .653 (Pearson) and 0155 [DF] and CEPA was only .482 (Pearson)
2007, 070 and CEPA was .699 (Pearson) and 0155 and CEPA was only .581 (Pearson)" (private emails, Sunday, Feb 17, 2008, 4:04 PM).
By 2009 even these posted correlations and results were no longer available on the college website. Fortunately, the researcher had printed out and retained a hard copy of the reported results and correlations performed by the college's Academic Services department for AY 2005/2006 before its subsequent disappearance. (See Appendix L for scanned copies of the KCA reports and the email exchange.)

Another bivariate analysis was performed on the MLR results of factors that were identified as potential performance indicators of academic success for the students of this cohort. This is further discussed in Sub-question three of Research Question (B) (pages 203-205).

It would appear, on the basis of the research findings, that there is reason to believe that strong predictive validity is possible for a regionally-developed English proficiency exam, like the CEPA. Even so, such findings must be treated with caution. The factors which influence the ability of an assessment instrument to predict are extremely complex and must not be over-simplified. The variables that may have the potential to positively affect academic performance, at least amongst this cohort, are discussed in the following section.

5.4.2 Research Question (B): What variables appear to positively correlate with students' ability to succeed academically?"

Bellingham (1993) referred to Graham's 1987 study when she noted that "diagnosis, prior to entry, of a potential student's proficiency in English provides valuable information from which the student and academic advisors can negotiate a pathway to academic success. Also, evidence is mounting to indicate that, especially at the lowest levels, language proficiency is a primary factor, along with individual differences, educational environment, and social context that influence academic outcomes". Since it was the researcher's intent to explore whether predictive validity can be more readily established, and with greater dependability, the more homogenous the population the test-takers hail from, it was important to determine the extent to which students who took the CEPA did share commonalities.

Sub-question one: Are the ages and academic levels of the research participants homogenous?

The overwhelming majority of the students who sat for the CEPA in this study were in the age range of 17-25 (95%). They were entirely male and Emirati (100% for each). Additionally, the vast majority of students indicated an unmarried marital status (94%), as well as remaining in their family's residence (96%).

As for the students' reported academic levels, it's not surprising that the majority of the students enrolled at this college graduated from public (government) high school (79% of the respondents), since enrollment to this college is restricted to Emirati and Persian Gulf nationals. The majority also reported that their CEPA scores were between 140 and 169.9 (65% of the respondents). While not actually an academic level, the number of college students who'd finished high school in the Arts stream was only slightly higher (53%) than those who'd finished in the Science stream (37%). As indicated by the students themselves, there was a high number of parents with little or no education: 44% of their fathers and 64% of their mothers were either illiterate or didn't finish school.

The results of these particular items, with the exception of the high school streams, would seem to indicate a high level of homogeneity of the test-taker population. Because of this, one would expect that calculations of the predictive validity of the CEPA to be more reliable than with more diverse populations, since the variables that affect students' success or lack of it may be assumed to be similar. Conversely, Bachman (2008: 278-279) referred to several studies in his consideration of the potential of diverse test taker populations to negatively affect language test results. The exploration of this issue formed part of the main thesis question: that high predictive validity can be achieved in populations of test-takers with very similar backgrounds and experiences.

Sub-question two: Is the socio-economic background of the participants homogenous?

Whilst not directly queried of the students in the questionnaire, a sense of their backgrounds may be gathered from their responses to items 7 and 8 about their parents' occupations, and items 10 and 11 about any second job parents have, in addition to the interview responses of the college counselor about this topic. As reported by students in the questionnaire, most of their mothers do not work, and most of the fathers are in lower-paid professional jobs, like policemen and civil servants. The college counselor was not able to able to provide any numeric data about the economic strata of students enrolled at the college, but he did point out that students coming from the Western Region were well-known to be disadvantaged in several ways: the poor quality of teaching and education there, the straitened economic circumstances of many of them, widespread illiteracy and high unemployment (not to mention the debilitating effects of a long commute back and forth to Abu Dhabi – not a socio-economic issue per se, but one which does tend to cause high absenteeism and consequently negatively affect students' ability to succeed).

It was expected by the researcher that students would report the sort of employment their parents have that they did report, but the college counselor's revelations about the differences in economic stature amongst Emirati students was a bit surprising, considering the general impression one gets of those who have lived in the UAE of a generally well-to-do national population. (See Appendix F: College Counselor interview, pages 2-3 and 4.) Even so, one would not expect a student body of almost 2000 to all belong to the same socio-economic stratum. Logically, one might acknowledge that a person's socio-economic status has an effect on a student's ability to succeed academically, but more research should be done before it's possible to qualify in any coherent way what differences it may or may not have. Again, the entire cohort shares a common culture (Among other things, they are all Emirati citizens, and all Arabic-speaking.), as well as a common religion (Whilst acknowledging that degrees of adherence differ amongst individuals in any society, these students do tend to share the same cultural interpretation of Islam.).

Even though the socio-economic homogeneity of the student cohort, within the confines of this study, cannot be determined with empirical clarity, they do appear to share a great deal in common with each other with regards to their social identity at least, if not economically as well.

Sub-question three: What are the students' motivations for, and attitudes towards, learning English?

The responses of the entire cohort to the final eight questionnaire items about their motivation for learning English were tabulated. (See Appendix D.) For both integrative and motivational items, the responses were overwhelmingly affirmative: the responses ranged from 92-99% agreement. The only anomalies, even though still very affirmative, were items #35 (87%), which was about wanting to know English to appreciate English art and literature, and #38 (74%), which was about learning English to gain more respect in the community.

Because of the exceptional uniformity of their responses, it was not possible to draw any viable, defensible conclusions about their motivations for learning English. For this reason, the researcher made a point to ask the volunteers in the students' group interview about the same issues. (See Appendix D: Transcribed Group Discussion One.) When queried directly about their motivation/s for learning English, a participant said that they need English for their future, and many others agreed. He also said that we are in the age of technology, and English is the recognised language of technology. Students also mentioned that English is the universally understood language in all the world's airports and banks. Another added that even here in their own country, officers in the Army must learn and use English. They mentioned that it is the only language through which they can converse with people from all nationalities. Another student said that in the UAE, you can't get much done if you can't communicate in English.

About their attitudes towards learning English, their responses were very positive. They felt and understood the need to be strong in English in order to communicate effectively with non-Arabs, and in order to get a good job. Even though every participant acknowledged that a sound knowledge of English is a requirement for the most lucrative sorts of employment for Emiratis, not all were happy with this state of affairs. For example, the group interview participants were asked a follow-on question about the emphasis placed on learning English in the UAE. (See Appendix D: Group Discussion Two.) One student articulated his dissatisfaction with the preeminence given to English as early as primary school: "Young students now have four English books and only one book for Arabic. Four! They are concentrating on English more than Arabic. This generation will not know Arabic the way they should. Why should they concentrate on English more than Arabic at the primary level? It is not a good thing. Young children in private schools no longer know Arabic!" He went on to say that children should be given a strong foundation in Arabic at the primary level and then given an improved education in English from the preparatory level onwards. The opinions voiced in the group interviews mirror those heard amongst Arabs in general. As mentioned earlier, one of the comments left by a student in the free response section at the end of the questionnaire was "This questionnaire did not ask about the reason for the lack of student interest in studying English". Counter to these individual student comments, most of those who work at the college, as revealed in the interviews with personnel from the college, felt that students were very keen to learn and even excel in English. (Appendix D.)

The effects of attitude on a student's academic progress is not only limited to the study of English, as mentioned by the DF Co-ordinator, "[I]t's not nice to be in a situation where you're failing. I mean they might be better if they were put in a situation or a context where they could excel, whatever area that might be - not academic, but something else. But to constantly be put in a situation where they're going to fail - I think it's very demoralizing. …It must have an effect".

Based on the discussions with the students themselves and their responses in the questionnaire, echoed by those who were interviewed, it is clear that their motivation for learning English is primarily instrumental, rather than an integrative: They need to be proficient in English in order to get a good job. This is not surprising considering that English has become the lingua franca of the UAE, and that most avenues for advancement and/or promotion in any field rely upon one's English competency.

Is it possible to identify variables that will enhance or ensure predictive validity?

A concerted effort was made to identify as many variables as possible that might affect students' academic success. This was done in order to identify key variables which could serve as performance indicators. The researcher initially included in the questionnaire variables that she suspected as being important, based on her interaction with students in the same demographic group over a period of years. The fine-tuning of this data collection instrument was further augmented by students themselves in the piloting of the questionnaire when the intent of the questionnaire was explained. However, even the most exhaustive and time-consuming attempt to consider and/or identify all the variables that may affect either academic success in general, or exam success specifically, will never be able to do so. This is because the whole process is extremely complex, and variables may range from an individual agonizing over one item in a test to a culturally-related perception held by the majority of the particular society. In short – it's an impossible task.

One issue was included in the protocols of the semi-structured interviews, and those interviewed were queried about it, but it hardly registered as having any importance: and that is the concept of 'cost', as investigated and explained by Bannerjee (2003: 93). She defined 'cost' as "the additional time and effort the students had to give to their studies, over and above the time and effort … that a native speaker of English might have to give [for the same work]. Additionally, if students had to lower their ambitions for their degree results [or indeed the course of study they wished to follow] because they were not coping with the linguistic demands of their courses, this was considered to be 'cost' [as well]" (ibid.). She refers to the damaging psychological 'cost' to individual students, as well as the larger issue of potential lost to society in general. Most of the interviewees acknowledged that this consideration was not a recognised part of any admissions, teaching or learning process. Perhaps not surprisingly, the only interviewee who recognised and stressed this as a serious and potentially destructive issue was the college counselor.

Variables that may correlate with student academic success: the MLR

Relevant to Research Question B are the results of the multiple linear regression, which compared students' first year final course results with their questionnaire responses in order to explore whether any of the variables students (and others) pointed out as important could predict or be used as indicators of future academic success.

The predictors identified by the MLR for the Diploma Foundations (DF) students' first year final marks for both English and Maths were the CEPA scores, their final high school (aka secondary) English and Maths marks respectively, as well as questionnaire responses regarding the nationality of their mothers (Emirati/Persian Gulf national), the college major they planned to enter (group A), and that their motivation to learn English was not mitigated by their desire to know English art and literature, but rather from a desire to garner greater respect from their community. English predictors also included what type of occupation they were aiming for and the fact that it took them less than 15 minutes on average to get to college from home. The predictors identified by the MLR for the Diploma Foundations students' first year final marks for Maths included other questionnaire responses about the number of siblings they had (less than 5), the type of high school they'd attended (private) and which 'stream' (science), a strong lack of satisfaction with their DF placement at the college and their college major preference (business). The predictors identified by the MLR for the Higher Diploma Foundations (HD) students' first year final marks for both English and Maths were students' CEPA marks, and the type of high school they'd attended (model or private) and which 'stream' (science or arts). The only other English final mark predictor for HD students turned out to be the educational level reached by their mothers. In contrast, other predictors for Maths final marks for HD were the students' GSC marks and their final high school Maths marks, the occupations of their fathers (group B – highly educated professionals) and the occupation they themselves desired (group C – skilled professionals).

Interestingly enough, almost none of the predictors revealed by the MLR were counter-intuitive. The CEPA was a predictor for all four analyses (Maths and English final marks for both DF and HD), which is not surprising given the results of the relatively strong (and/or statistically significant) correlations between the CEPA and students' final marks in high school English, and their final marks for their first year in college. The statistical significance of attending private school and being a graduate of the science stream is not surprising either. Having been a resident of the UAE for 20 years and taught in private high schools there for 12 years, the researcher has witnessed the typical emphasis and focus on Maths and sciences in schools, especially private ones, in the UAE. This is for several reasons. As Rumsey mentioned (in Davidson, Coombe and Jones (eds), 2005: 164), parents see this as value for money since it is widely believed that more lucrative careers are available to students who are strong in these subjects. Additionally, the Arts stream is both viewed as a stream for those who cannot manage the science stream and as a focus of study in which, as ESL or EFL learners, it is difficult for non-native speakers to excel.

It is important to note that all the question responses of the non-truncated sample were entered into the MLR to explore which of them might be identified as significant. One of the aims of the questionnaire was to assess whether variables that were identified as important for inclusion by students themselves (in the pilot of the questionnaire), and others directly involved with the students were actually statistically significant. Naturally, those that were not significant were automatically not included in the output of the statistical analysis itself.

These predictors were tested on part of the participant sample which was set aside specifically for this purpose. The predicted final English and Maths marks were compared to these students' actual final English and Maths marks using the Pearson's correlation coefficient. It was found that most of the predictors correlated remarkably well, as reported in the previous chapter (See pages 171-180).

What inferences might be drawn from this analysis? It has been possible, as revealed by the results of the MLR, to identify a few variables which appear to be positively connected with future academic performance, with at least one caveat: Although the statistical results look promising, they must be viewed with extreme caution because of the aforementioned issue of the impossibility of accounting for all potential variables. As Bachman (2004: 38) pointed out, even during the sitting of the exam itself, many subjective decisions, conscious and sub-conscious might be made by the test taker himself (different "subjective perspectives" about how to approach the test; "different subjective strategies for completing tasks"). There is no way to account for all such possibilities. Because of this subjectivity, variables identified as significant might be useful to augment the profiles of likely candidates, but if such variables are taken into consideration by the admissions office, they must not form the only, nor even a primary, part of the admissions process.

The focus of this study was not to concentrate on or investigate test design theory. The term "design" in this context means design choice. Likewise, exam development topics have been dealt with primarily at the level of choice – the decision to embark on such an endeavor, choice of exam type and the justification of these choices, as well as issues of validity.

The CEPA's Cronbach Alpha, the college's records, and the results of this study's correlations notwithstanding, this researcher has also recorded students' dissatisfaction with the examination, questions of data manipulation by the college administration, as well as skepticism voiced by members of the college's staff and administration. As noted by researchers of other published studies of locally or regionally developed examinations (Wall, Clapham and Alderson, 1991; Akoha, 1988), and by this researcher anecdotally (having been employed in tertiary institutions in different countries around the world) these issues are not uncommon. However, it does seem that high correlations have been achieved. This is the critical question that this study has tried to answer: What is it about the CEPA, or indeed the college where the research was undertaken or the students themselves, which accounts for this difference?

An exhaustive search of information about experiences similar to that of the development of the CEPA uncovered only five such efforts: In Benin, South India, Sri Lanka, Iran and Mexico. (See pages 64-72.) All but the project in Mexico reported less than satisfactory results. Let us weigh some mitigating issues. Much commonality may be observed in the issues reported. For instance, all the studies noted the local authorities' desire to either develop their own assessment instrument, or improve and/or validate the existing instrument, with the sole exception of the case in Iran. Resistance to fundamental and necessary change was noted in all studies, albeit to differing degrees. All the studies spoke of strained resources, as well. They all spoke of the lack of professional test development training (at least initially, e.g. the EXAVER project in Mexico) of the local test writers. All the authors spoke of at least the need for the validation of the instrument, if not the actual validation process. Major reported difficulties included most notably the lack of funds devoted to the effort, as well as training of local test developers which was poor or non-existent (Benin, South India, Sri Lanka, Iran). All of the articles mentioned issues of clashes in culture and test development sophistication between the local developers and either the consultants or standards considered the norm in so-called developed (i.e. UK, US) nations. These observations are represented in Table 5.1 on the following page. The developmental experience of the CEPA is also included. (The information in the table was gleaned from the cited articles in Section 2. of this study, and for the CEPA, from the semi-structured interview with the CEPA supervisor for this study, the researcher's experience as an item writer for the CEPA, a conference presentation on the

CEPA referred to below ("The CEPA: Language testing in the political arena", Dubai, UAE, 17 Nov. 2005) and an article by Brown and Jaquith in O'Sullivan (2011: 244-261).

"There are many testing situations which are under-resourced, both in human and material terms, where establishing test validity, especially for a new test, becomes even more of a challenge. [1] Opportunities for empirical pre-testing may be limited or non-existent. [2] Worries about test security mean that tests have to be constructed anew for each administration, without the possibility of calibrating items across versions of the same test for the same year, one of which will be selected by a person in high authority, so that the testers themselves cannot be accused of bias or corruption. [3] Gathering independent data on candidates' abilities can prove a logistical nightmare" (Wall, Clapham & Alderson, 1991: 209).

Table 5.1: Facts from published studies of regionally-developed English proficiency exams

	"Culture Clash" noted	**Strained resources**	**Training of test developers**	**Psychometric sophistication achieved**	**Validation sought**	**Validation achieved**
Benin	Yes	Yes	No	No	Yes	No
South India	Yes	Not addressed	No	No	No	No
Sri Lanka	Yes	Yes	Yes, by British consultants	Improved, but basic	Yes	Yes
Iran	No	Not addressed	No	Yes	Yes	Yes
Mexico	Yes	Possibly	Yes, by British consultants	Yes	Yes	Yes
UAE	No	No	Yes, by UK & US consultants	Yes	Yes	Yes

The decision to develop and expand the CEPA from its original purpose as a "low-stakes" placement tool to a high-stakes high school exit exam and tertiary admissions criteria, as Jaquith clarified in a presentation about the CEPA in the 2005 Current Trends in English Language Testing Conference ("The CEPA: Language testing in the political arena", Dubai, UAE, 17 Nov. 2005), was motivated by a desire to have a test which would not offend the cultural and religious sensitivities of the UAE people, and one which would address the specific language level requirements of the three national tertiary institutions of the UAE. Additionally, local development allowed the team to choose a test design most suited to the different technological and physical locales available. For example, the CEPA has occasionally been criticised for not including a speaking or listening element. In strictly practical terms, this is just not feasible. Consider that, right from the start, a huge additional investment in facilities and equipment (as well as maintenance personnel) would be necessary, in addition to the time, effort and again, expense, in developing a training module for the examiners. Add to this the fact that the pool of native speaker, non-national examiners is a very transient group, and CEPA could be faced with the potential nightmare of training enough examiners to test more than 20,000 students almost every year. Efficiency and economy was also at the heart of the decision to use bubble sheets for the answers, facilitating the scoring of a large number of students in a short time, an important consideration for the CEPA, whose thousands of marks must be recorded and uploaded to the Ministry of Higher Education within a few weeks of the exam, in order to make timely admissions decisions for students.

5.5 Implications for the design and development of regionally or locally produced proficiency exams

Major impediments to developing locally-produced exams are the substantial cost, time and effort involved. Any organisation wishing to pursue such an undertaking should be aware of the investment of time, resources and money such a project requires in order to produce an assessment which is psychometrically sound and which achieves high validity.

Advantages include the potential to be more culturally sensitive to students and addressing more specifically than internationally produced English proficiency instruments the particular linguistic needs (i.e. admittance to the national university or college) of regional stakeholders. (See Appendix M: CEPA sample exams contain items that are culturally specific to the UAE and the region.) Another advantage is the professional test-development experience gained by nationals involved in the process. This benefit of experience may be a major justification for the substantial investment required for the success of such a project. One observation may be that, having invested so much in the regional development of the exam, its primary appeal for regional stakeholders is one of a sense of ownership. This may indeed be a factor, but this does not negate the fact that, as explained by the interviewed CEPA supervisor, and as demonstrated in Brown and Jaquith (2011: 260), that the CEPA has established a high level of internal validity as well as consistent reliability.

There needs to be a collective awareness that recognises the serious and very real difficulties involved with encouraging and implementing change in language testing. The goal/s of the stakeholders must be clear and articulate, professional consultants can provide valuable assistance especially in the preliminary stages of the project, and the development process should be viewed as one which evolves and changes to meet new and changing needs. As

Akoha (1991: 204) pointed out, “Familiarity with experiences elsewhere and particularly with those in similarly difficult circumstances may shorten the march and prevent us having to reinvent the whole wheel”. Whilst recognising the challenges in local or regional test development, we do have the examples of the EXAVER (in Mexico) and CEPA (in the UAE) projects from which valuable lessons can and should be taken. This is further discussed in the following chapter.

5.6 The Thesis Question

In this study, we have thus far focused on the research questions, their corresponding sub-questions, and the findings. In this section, we shall address the core thesis question.

Is strong predictive validity possible for regionally-developed English proficiency examinations used for admissions and/or placement decisions?

As noted in section 2.13 (pages 65-68), predictive validity studies involving international English proficiency exams typically report comparatively low correlations between the exam scores and subsequent course achievement. The findings of this study seem to offer some vindication for the attainment of high predictive validity, but there are so many potentially confounding variables that this conclusion must be considered with a great deal of circumspection. As Bachman (2008: 31) pointed out, “When we design a test, we cannot incorporate all the possible factors that affect performance”. This is especially true in any attempt to apply this to the universality of experience.

The research questions were designed to answer the thesis question. Have they done so? We consider the thesis question and our intuition says “Yes”, but has the gathered empirical evidence supported this intuition? Yes, to a point because it is difficult to do so with any finality, and it is difficult to do so in a way that establishes a precedent with certainty. Additionally, the intention of case study research is not to discover any ultimate Truth or Reality, but to note, classify and verify phenomena for which replication is possible, thereby establishing a foundation of shared experiences that inform the academic community and hopefully affect change for the better.

5.7 Summary: Significance of the Findings

This was a predictive validity study – not an uncommon genre in the field of Applied Linguistics – but this one differs from others which precede it. It was undertaken partially because it does not suffer the inadequacies of most other predictive validity studies.

Most notably, the student cohort from which data were collected was not a truncated sample. For this particular academic year, all those recent high school graduates who wished to pursue their studies at this college were able to do so, regardless of their CEPA score. Admittedly, those whose scores were below the cut-off score for admission to either the Bachelors programme or the Higher Diploma programme were placed in the "lower" stream, but they were in fact granted admission and access to higher education. The researcher was given not only the entire cohort's admissions information, but also their final marks from their first year at college. Additionally, the questionnaire was completed by the entire 'Foundations' cohort for the academic year 2007/8. (The only exception was one HD section of approximately 20 students who were not brought to the auditorium to respond to the questionnaire.) Admittedly, it did become necessary to truncate the results: from the original 350 student questionnaires, only 225 were included in the comparative statistical analyses and the MLR owing to response papers in which the student chose not to identify himself, thereby making it impossible to compare responses with the data from the Registrar's office.

Another way in which this study differs from most other predictive validity studies is the overall homogeneity of the student population. This has been a bane of other predictive validity studies, making the attempt to identify variables that positively contribute to success (or negatively affect it) extremely difficult or even possible. Other research projects which have studied the predictive validity of a regionally-developed proficiency exam (i.e. the EXAVER project, Mexico) have also noted the advantage of having a relatively homogeneous student population. (See quote, page 70.)

This study holds specific significance for the college itself, as well, as it provides empirical proof of not only the success of the CEPA as a major component of admissions criteria, it also points to important issues to be addressed at the college level (as revealed in students' questionnaire responses and comments): the need to explore the provision of more night classes and a re-examination of the rule against students being employed whilst studying at the college, for example. Training to raise the awareness of instructors with regards to the development and preparation of progress tests and classroom assessments also seems indicated (noting the admission of the HD coordinator that he himself had developed the department's progress tests, and that these exact same tests had been repeatedly given for the past seven years).

Evidence gathered in this study also points to the possibility of identifying variables amongst a student population that has a very similar demographic profile – variables that correlate strongly and positively with academic success. With the exercise of great caution, it is possible that these variables could support not only the admissions process, but also be informative for recruiters and college counselors. For example, Diploma Foundations students' responses for two of the motivation items about learning English were identified by the MLR as highly predictive. Convincing conclusions also cannot be made on the basis of one year's findings. However, if such analyses were performed over a period of several years and certain patterns discerned, even though this information cannot be used to make admissions decisions by itself, it could be a factor to weigh in a candidate's favor along with his other qualifications.

In Chapter Five, an overview of the study summarised the focus of the research. In addition, and more importantly, the major findings of the research study were presented. Finally, reflections upon the significance of the findings of the study were brought forward. Chapter Six concludes the study with consideration of the implications of the findings, limitations noted and directions for future research suggested.

6: Conclusion

6.1 Introduction

In this chapter, we bring to a close the research study with a consideration of the implications of the research on theoretical perspectives, the implications for practise and observed limitations of the study. Implications for future research based on the findings (or indeed gaps discovered in the findings) of this study are then noted, followed by closing comments.

6.2 Implications for Theory

So what does the present predictive validity study contribute to the current state of knowledge – more disappointingly low correlations, another apologetic explanation of how the research findings are questionable as a result of using a truncated sample? No – actually the opposite, but does this now really vindicate predictive validity studies? Not really, but again, does this matter? Predictive validity is here with us to stay for the foreseeable future, in spite of its various shortcomings, at least because testing stakeholders who are not versed in, or even ready to be convinced of, its deficiencies still see it as a valuable quality for a language proficiency test to possess. And these deficiencies really have little to do with truncated samples.

> *"I have existed for years but very little has changed.*
>
> *I'm the tool of the government and industry too, for I am destined to rule and regulate you." (Zappa, 1973)*

There are a host of other problems that have the power to skew and invalidate results, not the least of which is the complex realignment of the whole educational system upon the introduction of a high stakes proficiency exam. One cannot help but recall the quoted lines above. This is what Broadfoot (2005), Christie (1995), and Shohamy (2000) referred to when they spoke of the effect of coaching and training of students to do well on the test and how this will *quid pro quo* affect the validity of the test as "a useful indicator of students' attainment" (Broadfoot, 2005: 131), and that "test scores are artificially inflated to the point of questionable validity" (Shohamy, 2000: 8). The dilemma of test coaching seems to (almost logically) coincide with exams becoming high-stakes assessments, or along with the introduction of a high-stakes assessment. The reservations of the DF supervisor with regards to coaching have been previously noted. However, in 2007 the CEPA was still relatively new as a high stakes assessment, its status having only changed a year earlier. In a specific question about coaching included in the student questionnaire, 79% of the responses indicated that they'd spent either more than 50% of English class time preparing for the CEPA. (Question item #29. See Appendix D.) In a separate question, 30% of the respondents indicated that they'd taken extra English classes in school to prepare for the CEPA, but an equal percentage (30%) indicated that they'd not taken any extra classes. (Question item #22. See Appendix D.) While it's impossible to say with any certainty what effect coaching may have had on the results, because of the newness of the assessment and the still rather unsophisticated preparation methods employed (recalling the comment of the group discussion member that they'd been required to memorize all 2000 words in an academic word list), it seems more likely that at least for this cohort, coaching would not have had such an effect as to majorly skew the research results.

There is a danger that the CEPA (or any regionally-produced proficiency exam), apart from its stated roles, has become or could become a "vehicle through which bureaucratic agendas [are] achieved" (Shohamy, 1997: 346). Shohamy (1997: 347), writing about English exams in Israel, described how the introduction of high stakes exams can "manipulate educational systems, control curricula, [etc., until] the testing policy becomes the *de facto* policy … through which control is exercised".

Another area this research has alluded to is that of fairness, specifically fairness as a quality by which to judge the validity of a test. While it is expected, as testing professionals, that everything in one's power is done to ensure that tests are as fair as possible, this cannot always be guaranteed. As Davies (2008) stated, "The pursuit of fairness is unnecessary on two grounds: first that it is unattainable, and second that it is unnecessary".

6.3 Implications for Practise

Admissions

While it is true that, within the confines of this research study and without access to the NAPO records of the validation process of the CEPA, it is not possible to ascertain or prove that the CEPA accurately assesses English language and provides correct estimations of competency with which to make admissions decisions at tertiary institutions in the UAE, it appears to be having some success at fulfilling its function as a gate-keeping and placement instrument. And it does this in spite of at least questionable and probably unethical practices of some of those entrusted with being the opposite; in spite of a myriad of complex and not wholly understood variables, it seems to work. This project continues to evolve also because of a dedicated team of professional test developers committed to an ongoing and iterative process of improvement. This study has not only confirmed that the CEPA is a predictor of performance at the foundations level, the MLR has also identified certain specific variables that seem to positively affect academic success. These variables differed according to the Foundations level and the subject for which the analysis was performed. These were presented and explained in the previous chapter (See pages 204-206.).

Implications for use of regional exams and internationally-developed exams

It is hoped that this research study will also have shed light on the problem of an over-reliance upon and complacency about the 'trustworthiness' of international proficiency examinations. For example, the stated purpose of the TOEFL test is to "evaluate the English proficiency of people whose native language is not English. ... The test has four sections: Listening - Measures ability to understand English as it is spoken in North America, Structure - Measures ability to recognise language that is appropriate for standard written English, Reading - Measures ability to understand non-technical reading matter, Writing - Measures ability to write in English on an assigned topic" (ETS, 1998). Even though the TOEFL is internationally recognised as a valuable assessment of the English level of foreign applicants to many western universities, the wording of the claims about it, as quoted above from the TOEFL catalogue, make it virtually impossible to ascertain whether TOEFL's claims about itself are valid or otherwise. First of all, there is no universally agreed-upon understanding of the concept of 'language proficiency' (Chalhoub-Deville, 1997: 4). Furthermore, only vague descriptions of the nature of the 'evaluation' are provided, with the exception of the Writing test, for which a more detailed scoring guide was provided. There are a few other 'loaded' terms in the wording of its claims which could be confusing, for example: what is 'native'?; what is "appropriate language'?; which reading matter is 'non-technical'?

A different and rather startling claim about the nature of this examination was also made: Examination questions "are then reviewed according to established ETS and TOEFL program procedures to *ensure that all possible versions of the computer-based test are free of cultural bias"* (ETS, 1998, italics mine). In the light of extensive research showing the impossibility of achieving this claim (Gipps and Murphy, 1994; Black, 1998), it is surprising that such an organisation could make it. Even when questions are intentionally designed to be completely de-contextualised, the prospect for bias remains in the actual choice of questions, question types and the style of language employed. This was never a claim made about the CEPA – actually quite the opposite. One of the recognised strengths of regionally-developed assessments is that it is possible to design items which are culturally familiar to regional test takers. A primary objective from the very beginning of the development of the CEPA was to intentionally be sensitive to the Islamic, Arab and Emirati cultures and traditions of its stakeholders.

The way we currently assess language competency is fraught with inefficiencies and inadequacies. This holds true whether we scrutinize well-known, internationally recognised language proficiency tests, or regionally-produced ones. Perhaps one way forward is for the worldwide testing community to emulate other fields in resisting and rejecting the globalisation of the instrument. What logic is there in using an internationally-produced proficiency exam for UAE secondary school students entering UAE universities? Surely locally or regionally produced examinations, supported by politicians ready to make a substantial investment in time and resources for this, and organised by testing professionals, have the potential to at least embody frames of reference which are not foreign to the local student population (as mentioned by those who authored the article about the EXAVER project in Mexico, and Nakamura, 1997: 3-21, who studied the implications of the existence of vague linguistic expectations in the IELTS writing module in Australia). The CEPA continues to be designed specifically for the Persian Gulf culture in general, and for Emirati nationals in particular with liberal references in various item stems and option clusters to traditional elements of Emirati life and names familiar to citizens of the region. (See Appendix M for examples from CEPA sample exams.) The author of this study was herself trained by, and did some work for CEPA as an item writer. Trainees were cautioned to avoid developing items whose content might distract Emirati students from the actual task at hand. Submitted test items were also reviewed by CEPA supervisors, who are themselves either Emirati nationals or long-time residents of the UAE.

International examinations have also been criticised for being expensive. Independent and government-sponsored research continues to search for an equitable solution to the resolution of the coursework/examination dilemma. The financial cost to test takers and other consumers is considerable. As described by Patricia Broadfoot (1984: 12-13), “"It is not surprising that considerable sums of public money have been available to national research institutes and psychometricians generally to develop and evaluate different assessment techniques. Research and development of this kind has grown from being a scholarly pursuit in the early years of this century, into a major business..." As demonstrated in the article about the EXAVER project in Mexico, the regional development of proficiency exams, whilst necessitating a large initial investment, eventually pays for itself in many ways, not the least of which is the training of nationals in state-of-the-art test development. In this way, a sense of pride and ownership may be expected to be achieved, which does not necessarily imply a less valid instrument.

6.4 Limitations

The veracity and strength of the inferences one makes based on the outcomes of research validates the research study itself. What detracts from the strength of these inferences is what Teddlie and Tashakkori (2009: 299) describe as “threats to inference quality”. They list eight possible limitations which may restrict or negatively affect the inferences which have been made. Seven of these eight ‘threats’ will be considered here in the context of the inferences made in this research study. Their category of “pretesting” was left out because it is not relevant to this particular study. (All the quotes in this section, unless otherwise noted, are from Teddlie and Tashakkori, 2009: 299.)

Selection

The first category of such threats to inference quality is that of selection. This refers to the possibility that “certain attributes of one group are different from another before the study starts”. Whilst a distinctive feature of this study is the homogeneity of the test taker cohort in their backgrounds, culture, language and religion, undoubtedly differences do exist. Most apparent of these is social class. Those from the ruling families and those from wealthy families may enjoy advantages over those who are less fortunate. Other noted variables have been distance from college and differences in school experiences (i.e. private/public). It would be difficult however to determine (quantify) how such differences might affect or skew the results of this study.

Another ‘selection’ issue has to do with the cohort itself. This study only examined the strength of CEPA as a predictor for the foundations courses at the college. The researcher was not granted access to first year matriculated Bachelors students. Referring to the observation made at the beginning of this report about the claims made (or often expected of) English proficiency exams to not only gauge language ability, but also to be a factor in the predictions educational institutions make about students' ability to succeed academically, it would have been beneficial to have GPA scores for the students, but GPAs are not calculated for (pre-Bachelors) Foundations students. This would also facilitate comparison of these results to other studies.

Finally, with regards to selection, there is the issue of 'self-selection', which refers to those students who chose not to provide their names on the questionnaire, thereby making it impossible to include them in the comparative analyses. This is discussed in the 'Attrition' section below.

History

Apart from the primary focus of this limitation, which is a differing background history of two groups being studied, "history" can also include events happening to a group of individuals beyond the event that the researcher is studying". For this study, this would be what has already been alluded to: i.e. that it is simply not possible to account for all the possible variables which might affect, positively or negatively, a person's academic standing. That is beyond the event which was studied.

Statistical Regression

There is the problem of an established "extreme" case or group. The cohort of students which was the focus of this study was an extremely rare find: a sample that was not truncated, and as such would be a difficult sort of study to replicate. In addition, while the homogeneity shared by the members of the cohort would be extremely difficult to find in more culturally diverse societies for the purpose of study replication, there still exists large numbers of Asian and African societies about which this would not be unusual.

Maturation

Here, Teddlie and Tashakkori refer to a pre-test/post-test research situation, where the natural maturation of the participants is a factor, and not just the independent variable. Although this was not a part of this study, it would be illogical to assume that the varying degrees of maturation, both physical and psychological, of the students in this study had no effect on their final academic outcomes at the end of the semester, a full year after they took the CEPA.

Instrumentation

This refers to the possibility that the differences and/or similarities between the CEPA scores and the students' final marks at the end of their first year could be the result of random measures, which would be very difficult, if not impossible, to account for. The college claims high internal validity statistics for the KCA (final exams), but what of the various assessments the students take during their college year from different instructors, very little of which is monitored by the administration? The college has acknowledged this issue, and that is the rationale for allowing the teachers' more subjective input only a small percentage of a student's mark. Nevertheless, departmentally-prepared assessments are not immune from serious drawbacks and/or concerns, since these are also not centrally monitored. One recalls the revelation of the Higher Diploma Co-ordinator: "...the progress tests… Well, I wrote them and we administer them. ...[The progress tests] mirror the final exam completely. We familiarise students with the… we teach to the test. … We've used the same exams for about 8 years. We've polished them. … And they work." With unaccounted-for variables such as this, it can pose questions as to the veracity (and replicability) of the results attained. Fortunately for this study, 070 English departmentally-prepared assessments for Higher Diploma Foundations represent a very small percentage of the total quantitative data collected. Even so, such anomalies must be acknowledged. This can help guide future studies in the college itself, and clarify issues for researchers wishing to replicate this research elsewhere.

Implementation

As with any study which includes qualitative data collection strategies, "The obtained relationship between variables might be a result of researcher expectancy or participant reactivity to being studied". While perceptions are important, they may be coloured by enthusiasm, suspicion, vested interests, a lack of comprehension and/or recollection, etc., and may not reflect actual relationships as they exist. The researcher felt this was particularly true with regards to the results of the eight motivation questions added to the end of the questionnaire. By all accounts – the students themselves, the administrators and instructors – the students' primary reason, and most pressing motivation, for learning English was to get a more lucrative job. The researcher felt that the students did not really take the motivation questions seriously, since the overwhelming majority expressed strong agreement for ALL the motivation questions. Participant reactivity to being studied may very likely have been a variable in not only how students responded to the questionnaire, but also in their discussion responses, and even in the pilot of the questionnaire itself.

Another aspect of "participant reactivity" noted in the data collection phase was fear of reprisal or backlash regarding any information volunteered. This was a major issue with regards to finding participants for the group discussion who were willing to not only sign a waiver but also to be recorded, even though assurances were made about the protection of their identities. Additionally, the college counselor actually requested that a significant part of his interview response be struck from the record for his explicit concern of reprisal resulting from his expression of candid opinions in the interview itself. Fear of reprisal may also have played a part in many of the responses in the qualitative data collected.

Attrition

Although some students (17 of them) did withdraw from the college between the administration of the questionnaire and the final exams, the size of the cohort remained large enough not to jeopardise or skew the results on that basis alone. (From the original 347 who took the questionnaire, 225 were included in the final analyses.) Another form of attrition which must be noted as a limitation of the study is the fact that of the 347 students to whom the questionnaire was administered, 122 of them chose not to write their names. Whether consciously or not, by doing so, they self-selected themselves out of the rest of the study. Self-selection is a limitation of qualitative research.

6.5 Implications for Future Research

The outcomes of research can broaden understanding in the field of study, but they almost inevitably inspire new questions that can and should lead to further research, where they may have broader impact and applications. The current research investigation has perhaps raised more questions that it has answered, about issues which, although did not fall within the focus of this study, are nonetheless important.

One important area that deserves consideration is the previously mentioned issue of Bannerjee's (2003) idea of the psychological 'cost' to ESL or EFL students who struggle with the linguistic demands of programmes where the medium of instruction is English.

The rationale and justification for the changes made in the roles and uses of the CEPA in 2006 and 2007 are not altogether clear, nor if there has been any validation effort for these instituted changes. Along the same line, the rationale for the calculation of the PI should be in the public arena. This information is important not only to the test-takers and other stakeholders directly affected by these decisions, but also to any other social scientist who wishes to benefit and learn from this apparently successful project. Another critical issue to set straight is obviously the glaring inconsistencies, gaps and changes in the data reporting of the college's administration.

The 'truncated sample issue' should now be considered a non-issue. This research has shown that even with a *non-truncated* sample, there remain irresolvable difficulties with selection (as mentioned on page) and with accounting for all the variables that may affect performance. This was examined in this study in the section about the 'Affection Domain of Test-Takers' in Chapter 2 (see pages 62-64). Difficulties with regards to test transparency and other ethical issues also have the potential to affect results, as noted by Davies (1997), Alderson et al. (1995), Shohamy (2001), Spolsky (1997) and others.

Finally, it would be interesting to further scientifically investigate the differences in the quality of education in the UAE between private and public schools in order to have empirical evidence from which informed conclusions may be drawn. The general assumption is that private education is 'better' than public. This difference amongst students was selected by the multiple linear regression as a statistically significant predictor of future academic success of students at HCT. This warrants further investigation.

It has been stated that "because they are intensive, [case studies] bring to light important variables, processes and interactions that deserve more extensive attention". Further research on the CEPA is important to support or negate the findings of this study, and to provide more thorough information to others who may wish to pursue the development of region-specific proficiency exams. Another strength of the case-study, noted by Stanovich and Stanovich (2010), is that they pioneer new ground and often are the source of fruitful hypotheses for further study. One of the hopes of the present research findings is that it might be replicated in other areas where a homogenous student body of test takers could potentially be better served by the development of locally produced proficiency exams, thereby creating a pool of information from which might be drawn more useful and stronger conclusions.

6.5 Summary

There are an almost infinitesimal number of variables that may factor into a student's academic success, or lack of it. Some of these variables are connected with educational testing. Examining the ethics of testing has been a depressing and disheartening avenue of exploration. Liz Hamp-Lyons recently echoed this sentiment in a thread on the L-testL forum (27 Nov. 2011) when she said, "I find the discussion about values depressingly knotty to unpack". So much of what we do as testing professionals centres around numerical equivalents. It is very easy to get lulled into a false state of psychometric complacency (or even superiority) when the numbers coalesce so well. Academics like Broadfoot, Shohamy, Davies and Alderson have over and over raised a flag of caution about this, reminding us that testing is, at its heart, a very human enterprise. It is important to acknowledge the limitations of our statistical, pseudo-perfection and return to the core of why we got involved in this area of expertise to begin with. For most of us, we are and have been guided by humanistic ideals of doing our best to help society come to terms with and ethically use instruments that have such a critical impact on every facet of education and individual success. It is not only our duty as testers to do good; we must also endeavour to do no harm, as much as is it within our capabilities to do so. And so, edging away from despairing of any major reform in how we measure students' linguistic abilities, we at least can glean from collections of research those kernels of what seems to offer a better way of doing things. Whilst speaking of action research in particular, what McDonough and McDonough (1997:33) noted may be applied to the research of case studies as well: that one of the goals of humanistic research should be to "… challenge entrenched structures of power and authority, to subvert autocratic and top-down decision-making procedures, and 'to [emancipate] individuals from the domination of unexamined assumptions embodied in the status quo".

BIBLIOGRAPHY

Akoha, Joseph (1988) "Curriculum Innovation and Examination Reform in Benin, West Africa" in Alderson, Charles & North, Brian (1991) Language Testing in the 1990s: The Communicative Legacy, Modern English Publications & The British Council, 198-208

Al Ghazali, Fawzi (2008) "Common Educational Proficiency Assessment (CEPA) Retrospective and Prospective Views", (PowerPoint presentation, CTELT 2008) http://usir.salford.ac.uk/21012/1/CEPA_Retrospective_and_Prospective_Views.pdf; last accessed July 3, 2013

Alderson, Charles & North, Brian (1991) Language Testing in the 1990s: The Communicative Legacy, Modern English Publications & The British Council

Alderson, Charles, Clapham, C., Steel, D. (1997) "Metalinguistic knowledge, language aptitude and language proficiency", *Language Teaching Research* 1, 2: 93-121

Alderson, Charles, Clapham, C. & Wall, D. (1995) Language Test Construction and Evaluation, CUP, Cambridge

Aski, Janice M. (1998) "Theory into Practice: Italian Quizzes and Exams as a Reflection of the Curriculum", *Italica*, Volume 75, Number 4, 477-491

Bachman, Lyle and Palmer, Adrian (1996) Language Testing in Practice, OUP

Bachman, Lyle (1990) Fundamental Considerations in Language Testing, OUP

Bachman, Lyle (2004) Statistical Analyses for Language Assessment, CUP

Bailey, Kathleen (1998) Learning About Language Assessment: Dilemmas, Decisions, and Directions, Heinle and Heinle Publishers

Banerjee, Jayanti (2003) Interpreting and Using Proficiency Test Scores, unpublished doctoral thesis, Lancaster University

Barkaoui, Khaled (2010) Language Assessment Quarterly; Vol 7, Issue 1, 2010 pages 54-74

Barrow, Robin and Milburn, Geoffrey (1986) *A Critical Dictionary of Educational Concepts*, Wheatsheaf Publishers

Bell, Judith (1993, 2nd ed) Doing Your Research Project: A Guide for First-time Researchers in Education and Social Science, Open University

Biggs, John (1998) "Assessment and Classroom Learning: a role for summative assessment?", *Assessment in Education: Principles, Policy and Practice*, Vol. 5, n 1 (Mar 1998)

Black, Paul (1998) Testing: Friend or Foe? The Theory and Practice of Assessment and Testing, The Falmer Press

Black, Paul and William, Dylan (1998) "Assessment and Classroom Learning", *Assessment in Education: Principles, Policy and Practice*, Vol. 5, n 1 (Mar 1998)

Boyd, Kenneth and Davies, Alan (2002) "Doctors' orders for language testers: the origin and purpose of ethical codes", *Language Testing*, 19 (3), 296-322

Broadfoot, Patricia (1984) Selection, certification, and control: social issues in educational assessment, The Falmer Press

Broadfoot, Patricia (1986) " Alternative to Public Examinations", in Nuttall, Desmond (ed.) Assessing Educational Achievement, The Falmer Press. 54-80

Broadfoot, Patricia (April 1996). Education, Assessment and Society: A Sociological Analysis (Assessing Assessment). Open University Press

Broadfoot, Patricia (2005) "Dark alleys and blind bends: testing the language of learning", *Language Testing* 22 (2) 123–141

Brown, J. D. (1990) "Where do tests fit into language programs?" *JALT Journal, 12* (1), 121-140

Carroll, Brendan (1991) "Resistance to Change", in Alderson, Charles & North, Brian (1991) Language Testing in the 1990s: The Communicative Legacy, Modern English Publications & The British Council, 22-27

Chalhoub-Deville, M. (1997) "Theoretical models, operational frameworks and test construction", *Language Testing* 14, p. 3-22

Cook, Vivian (1996, 2nd ed.) Second Language Learning and Language Teaching, Arnold: London

Cook, Vivian (2007) http://homepage.ntlworld.com/vivian.c/SLA/MotTest.htm, last accessed 29 July, 2007

Cotton, Fiona and Conrow, Frank (1995) "An investigation of the predictive validity of IELTS amongst a group of international students studying at the University of Tasmania". *IELTS Research Reports* (Vol. 1). Canberra: IELTS Australia. (p. 72-115)

Cope, Nicholas (2011) "Evaluating Locally-developed Language Testing: A Predictive Study of 'Direct Entry' Language Programs at an Australian University". *Australian Review of Applied Linguistics*, Vol. 34, No. 1, 40-59

Cresswell, John and Plano Clark, Vicki (2011, 2nd ed) Designing and Conducting Mixed Methods Research, Sage Publishers: LA, London

Cronbach, L. J. (1969) "Validation of educational measures" *Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton*, NJ: Educational Testing Service, 35-52.

Cronbach, Lee and Meehl, Paul (1955) "Construct Validity in Psychological Tests", *Psychological Bulletin*, 52, 281-302

Davidson, P. & Dalton, D. (2003). Multiple-measures assessment: Using 'visas' to assess students' achievement of learning outcomes. In Coombe, C.A. & Hubley, N. (Eds). *Case Studies in TESOL Practice Series*: *Assessment Practices*, Virginia: TESOL, 121-134

Davidson, F. and Lynch, B. (2002) Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications, Yale University Press

Davies, Alan (1988) "Operationalising uncertainty in language testing: an argument in favour of content validity", *Language Testing*, 5 (1), 32-48

Davies, Alan (1990) Principles of Language Testing. Oxford: Blackwell

Davies, Alan (1997) "Demands of being professional in language testing", *Language Testing*, 14 (3) 328-339

Davies, Alan (1997) "Introduction: The limits of ethics in language testing", *Language Testing*, Vol. 14, No. 3, 235-241

Davies, Alan (2003) "Three Heresies of Language Testing Research", *Language Testing*, Vol. 20, No. 4, 355-368

Davies, Alan (2008) "Does test validity depend on fairness?", paper presented at annual Language Testing Forum, Newcastle, UK

Denvir, Brenda (1988) “What Are We Assessing in Mathematics and What Are We Assessing For?” in Pimm, David (ed.) Mathematics, Teachers and Children, London: Hodder & Stoughton. 129-140

Denzin, N. K. (1988) The Research Act: A Theoretical Introduction to Sociological Methods. 3rd Ed. Prentice-Hall: Englewood Cliffs, NJ

Alan C. Eliott and Wayne A. Woodward (2007) Statistical Analysis Quick Reference Guidebook: With SPSS Examples, Sage Publications

Feast, Vicki (2002) “The impact of IELTS scores on performance at university”, *International Education Journal*, *3* (4), 70-85

Field, Andy (2005, 2nd Ed.) Discovering Statistics Using SPSS, Sage Publications; London

Florescano, A., O’Sullivan, B., Chavez, C., Ryan, D., Lara, E., Martinez, L., Macias, M., Hart, M., Grounds, P., Ryan, P., Dunne, R., Barradas, T. (2011), in O’Sullivan, Barry (ed.) O’Sullivan, Barry (2011) Language Testing: Theories and Practices , Palgrave, 228-243

Fox, Janna (2004) “Test Decisions over time: Tracking Validity”, *Language Testing*, Vol. 21, No. 4, 437-465

Fulcher, Glenn (1997) “An English language placement test: issues in reliability and validity”, *Language Testing*, Vol. 14, No. 2, 113-139

Fulcher, Glenn & Davidson, Fred (2007) Language Testing and Assessment, Routledge: Abingdon

Gardner, R. (1985) Social Psychology and Second Language Learning, Arnold, London

Gipps, Caroline and Patricia Murphy (1994) A Fair Test? Assessment, Achievement and Equality, Open University Press: Buckingham

Griffin, Patrick (2001) “Establishing meaningful language test scores for selection and placement”, in C. Elder, A. Brown, K.Grove, E. Hill, N. Iwashita, T. Lumley, T. McNamara, and K. O’Loughlin (eds.) Experimenting with Uncertainty: Essays in honour of Alan Davies, CUP, 97-107

Gronlund, NE (1976) Measurement and Evaluation in Teaching (3rd edn.). New York: Macmillan Publishing.

Guion, R. M. (1977). “Content validity–The source of my discontent”, Applied Psychological Measurement

Hamp-Lyons, Liz (1997) “Washback, Impact and Validity: Ethical Concerns”, *Language Testing*, Vol. 14, No. 3, 295-303

Hamp-Lyons, Liz (2013) From: Language Testing Research and Practice [mailto:LTEST-L@LISTS.PSU.EDU] On Behalf Of Liz Hamp-Lyons, Sent: 24 June 2013 11:34,To: LTEST-L@LISTS.PSU.EDU, Subject: Re: [LTEST-L] proficiency vs. competency

Harrison, A. (1983) A Language Testing Handbook, London: Macmillan Press

Hawthorne, Lesleyanne (1997) “The political dimension of English language testing in Australia”, *Language Testing* 1997; 14; 248-260

Hughes, Arthur (2003, 2nd ed.) Testing for Language Teachers, CUP, Cambridge

Hymes, D.H. (1966). "Two types of linguistic relativity". In Bright, W. *Sociolinguistics*. The Hague: Mouton. pp. 114–158

Kane, Michael (2004) “Certification Testing as an Illustration of Argument-Based Validation”, *Measurement: Interdisciplinary Research and Perspective,* 2:3, 135-170

Kane, M. T. (2006). "Validation", in R. L. Brennan (Ed.), Educational measurement (4th edition), Westport, CT: Praeger, 18-64

Kane, Michael (2008) "Terminology, Emphasis, and Utility in Validation", *Educational Researcher* 2008; 37; 76-82

Kerstjens, Mary and Nery, Caryn (2000) *Predictive validity in the IELTS test: A study of the relationship between IELTS scores and students' subsequent academic performance*. IELTS Research Reports (Vol. 3). Canberra: IELTS Australia. 85-108

Klein, Joseph & Wasserstein-Warnet, Marc (2000) "Predictive Validity of the Locus of Control Test in Selection of School Administrators", *Journal of Educational Administration*, Vol. 38, No. 1, 7-24

Klenowski, Valentina (1996) "Connecting Assessment and Learning", paper presented at the BERA Conference, University of Lancaster, Sept. 1996

Lee, Young-Ju and Greene, Jennifer (2007) The Predictive Validity of an ESL Placement Test. *Journal of Mixed Methods Research*, Vol. 1 (4), October 2007, 366-389

Leung, Constant (2005). "Convivial communication: recontextualizing communicative competence". *International Journal of Applied Linguistics* 15 (2): 119–144. ISSN 0802-6106. Retrieved June 20, 2013

Lim, Gad (2009) Prompt and Rater Effects in Second Language Writing Performance Assessment, doctoral dissertation for University of Michigan

Linn, Robert L. (1982) "Admissions Testing on Trial", *American Psychologist*, Vol 37 (3), Mar 1982, 279-291

Lissitz, Robert and Samuelsen, Karen (2007) "A Suggested Change in Terminology and Emphasis Regarding Validity and Education", *Educational Researcher* 2007; 36; 437-448

Lissitz, Robert and Samuelsen, Karen (2007) "Further Clarification Regarding Validity and Education", *Educational Researcher* 2007; 36; 482-484

Lyman, Howard (1991) Test Scores and What They Mean, Prentice Hall

Lynch, Brian K. (1997) "In search of the ethical test", *Language Testing*, Vol. 14, No. 3, 315-327

Mark, Jonathan and Goldberg, Michael A. (2001). Multiple Regression Analysis and Mass Assessment: A Review of the Issues. *The Appraisal Journal*, Jan. 89–109

Markus, Keith (1998) "Validity, Facts, and Values sans Closure: Reply to Messick, Reckase, Moss, & Zimmerman", *Social Indicators Research* 45: 73–82, 1998. Kluwer Academic Publishers

McDonough, Jo and Steven McDonough (1997) Research Methods for English Language Teachers, Arnold: London

Merrylees, B. and McDowell, C. (1999), "An investigation of Speaking Test reliability with particular reference to the Speaking Test format and candidate/examiner discourse produced", in *IELTS Research Reports Vol 2*, ed R Tulloh, IELTS Australia, Canberra, 1-35

Merriam, Sharan B. (1998) Qualitative Research and Case Study Applications in Education, Jossey-Bass: San Francisco

Metais, Joanna and Tabberer, Ralph (1997) "Why Different Countries Do Better: Evidence From Examining Curriculum and Assessment Frameworks in 16 Countries", *International Electronic Journal for Leadership in Learning*, 1(3), http://www.ucalgary.ca/iejll/metais_tabberer (Last accessed 31 July, 2009)

McNamara, Tim (2001) “Language Assessment as Social Practice: Challenges for Research”, Language Testing 2001; 18; 333-349

McNamara, Tim (2000) Language Testing, OUP

Mead, R.J. (2008) A Rasch primer: the measurement theory of Georg Rasch. Psychometrics services research memorandum 2008–001. Maple Grove, MN: Data Recognition Corporation

Messick, Samuel (1989) “”Validity”, in R.L. Linn (ed.), Educational Measurement, Macmillan, NY, 13-103

Messick, Samuel (1998) “Test Validity: A Matter of Consequence”, *Social Indicators Research* 45: 35-44

Moss, Pamela A. (2007) “Reconstructing Validity”, *Educational Researcher* 2007; 36; 470-6

Nakamura, Yuji (1997) “Involving Factors of Fairness in Language Testing”, *Journal of Communication Studies*, Sep 1997, n7; 3-21

North, Brendan (1991)” “, in Alderson, Charles & North, Brian (1991) Language Testing in the 1990s: The Communicative Legacy, Modern English Publications & The British Council

Nunan, David (1992) Research Methods in Language Learning, CUP

Nuttall, Desmond (ed.) Editorial in D.L. Nuttall (1986) Assessing Educational Achievement, Falmer Press, 1-4

O’Sullivan, Barry (2002) “Investigating variability in a test of second language writing ability”, *Cambridge Research Notes*, Issue 7, Feb 2002, 14-17

O’Sullivan, Barry (2011) “Language Testing”, in Simpson, J (2011) The Routledge Handbook of Applied Linguistics, p. 259-271, Routledge

O’Sullivan, Barry (2011) Language Testing: Theories and Practices , Palgrave

Osterlind, Steven J. (1997) Constructing Test Items: Multiple Choice, Constructed Response, Performance and Other Formats, Springer Publications

Page, G.T., Thomas, J.B., Marshall, A.R. (eds) (1977) The International Dictionary of Education, Nichols Publishing

Patton, Michael Quinn (1990) Qualitative Evaluation and Research Methods, 2nd Edition, Sage Publications

Porter, D. and O’Sullivan, B. (1999) “The effect of audience age on measured written performance”, *System*, Volume 27, Issue 1, March 1999, 65-77

Purpura, James E. (2004) Assessing Grammar, CUP

Rumsey, Laila (2005) “A Case Study of Comparative Assessment”, in Davidson, Coombe and Jones (eds) (2005) Assessment in the Arab World, TESOL Arabia publication, Dubai, p. 149-172

Sadler (1989) “Formative Assessment and the Design of Instructional Systems”, *Instructional Science*, 18:119-144 (Kluwer Academic Publishers)

Shaw, S. and Weir, C. (2007) Examining Writing: research and practice in assessing second language writing, CUP, Cambridge ESOL

Shepard, Lorrie (1993) “Evaluating Test Validity”, in L. Darling-Hammon (Ed.) *Review of Research in Education*, Vol. 19, Washington, DC, AERA, 405-450

Shiotsu, Toshihiko and Weir, Cyril (2007) "The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance", *Language Testing* 24 (1), 99-128

Shohamy, Elena (1997) "Testing methods, testing consequences: are they ethical? Are they fair?", *Language Testing* 14 (3): p. 340-349

Shohamy, Elena (2000) "Using language tests for upgrading knowledge: The phenomenon, source and consequences", *Hong Kong Journal of Applied Linguistics*, v5 n1 p1-18 Oct 2000

Shohamy, Elena (2001) The Power of Tests: A Critical Perspective on the Uses of Language Tests, Harlow: Pearson Education

Shohamy, Elena (2003) "Implications of Language Education Policies for Language Study in Schools and Universities", *The Modern Language Journal*, 87 ii, 277-296

Short, Deborah (1991) "Integrating language and content instruction: Strategies and techniques" Washington, DC: *National Clearinghouse for Bilingual Education*.

Sireci, Stephen G. (2007) "On Validity Theory and Test Validation", *Educational Researcher* 2007; 36; 477-481

Spolsky, Bernard (1997) "The ethics of gatekeeping tests: What have we learned in a hundred years?", *Language Testing*, Vol. 14, No. 3, 242-247

Stanovich, Paula and Keith Stanovich, (2010) http://www.nichd.nih.gov/publications/pubs/using_research_stanovich.cfm (last accessed 1/10/10)

Stansfield Charles (2008) "Lecture: 'Where we have been and where we should go'", *Language Testing*, Vol. 25, No. 3, 311-326

Stenhouse, L. (1998) "Case Study Methods", in Keeves, John P (ed.) Educational Research Methodology and Measurement: an International Handbook, Pergamon Press: Oxford, p. 49-53

Sternberg, Robert (2006) *"The Rainbow Project: Enhancing the SAT through Assessments of Analytical, Practical and Creative Skills"*, Intelligence, Jul-Aug 2006, Vol. 34, 4, p. 321-350

Stewart, D.W. & P.N. Shamdasani (1990). Focus Groups: Theory and Practice, Sage Publications: Newbury Park, London, New Delhi

Stobart, Gordon (2008) Testing Times: The Uses and Abuses of Assessment, Routledge, UK

Sutton, Ruth (1991) Assessment: A Framework for Teachers, NFER-Nelson: Windsor, UK

Talbot, Robert and Briggs, Derek (2007) "Does Theory Drive the Items or Do Items Drive the Theory?", *Measurement: Interdisciplinary Research and Perspective*, 5:2, 205-208

Tanner, David E. (2003) *"Admissions and Placement Testing: Enough is Enough!"*, US Dept. of Education Publication

Teddlie and Tashakkori (2009)

Weir, Cyril (2005) Language Testing and Validation: An Evidence-Based Approach, Palgrave MacMillan

Weir, Cyril (2008) "A cognitive processing approach towards defining reading comprehension", *Cambridge ESOL: Research Notes*, Issue 31, February 2008

Wiggins, Grant (1997) "Practising What We Preach in Designing Authentic Assessments", Educational Leadership, Dec. 1996/ Jan. 1997

Winn, Ralph B. (1959) *John Dewey: Dictionary of Education*, Philosophical Library (Pub.)

Wood, Robert (1993) Assessment and Testing, CUP: Cambridge

Yin, R. K. (2003, 3rd ed.) *Case study research, design and methods*, Sage Publications: Newbury Park

Yin, Robert (2011) http://www.scribd.com/doc/37102046/Robert-Yin-Case-Study-Research; last accessed 8/18/2011).

Young, J. W. (2001) "Differential validity and prediction: Race and sex differences in College Admissions Testing, Zwick, Rebecca (ed.) (2001) Rethinking the SAT, Routledge Falmer: NY, p. 289-302

Zappa, Frank (1973) "I'm the Slime", from the album *Over-Nite Sensation*, Montana Records

Appendices on Disk