EFFECTIVE COLLEGE TEACHING AND STUDENTS' RATINGS OF TEACHERS: WHAT STUDENTS THINK, WHAT FACULTY BELIEVE, AND HAT ACTUAL RATINGS SHOW IMPLICATIONS FOR POLICY AND PRACTICE IN TEACHING QUALITY ASSURANCE AND CONTROL IN HIGHER EDUCATION IN OMAN

AL-HINAI, NASSER,SAID

EFFECTIVE COLLEGE TEACHING AND STUDENTS' RATINGS OF TEACHERS: WHAT STUDENTS THINK, WHAT FACULTY BELIEVE, AND WHAT ACTUAL RATINGS SHOW

IMPLICATIONS FOR POLICY AND PRACTICE IN TEACHING QUALITY ASSURANCE AND CONTROL IN HIGHER EDUCATION IN OMAN

A Thesis Submitted to the University of Durham
For the degree of

DOCTOR OF PHILOSOPHY

By

NASSER SAID AL-HINAI

University of Durham
School of Education
United Kingdom

January 2011

ABSTRACT

EFFECTIVE COLLEGE TEACHING AND STUDENTS' RATINGS OF TEACHERS: WHAT STUDENTS THINK, WHAT FACULTY BELIEVE, AND WHAT ACTUAL RATINGS SHOW

IMPLICATIONS FOR POLICY AND PRACTICE IN TEACHING QUALITY ASSURANCE AND CONTROL IN HIGHER EDUCATION IN OMAN

Nasser Said Al-Hinai

This study examines the extent to which teachers' (N=248) and students' (N=968) perceptions of effective teaching and students' evaluations of teachers in six colleges of technology in Oman match or mismatch. It also investigates Omani students' (N=922) ability to identify the teaching dimensions underlying a widely used American instrument used for collecting students' evaluations of teachers and the extent to which the teaching dimensions found in Oman are similar to or different from those found in America and elsewhere in the West. In addition, the present research assesses the reliability of students' ratings in Oman and the effect of a number of course, teacher, and student background characteristics on these ratings.

Results showed that while teachers and students matched in their perceptions of various characteristics of effective teaching, they significantly differed in their valuation of many criteria of effective teaching. Differences were also observed between the two groups' perceptions of the validity and utility of students' ratings and the role of the student as an evaluator of teaching.

The results also showed that Omani students are capable of identifying most of the teaching dimensions underlying the standardised American rating instrument. A few factors, however, appear to be inseparable in the Omani context. The inter-rater reliability of students' ratings collected from Oman was analysed and found to be of good standard and only slightly lower than what was found in North America and Australia for the same instrument. Consistent with previous research, it appears, however, that students' ratings are affected by various student, teacher, and course background characteristics.

The evidence on the differences between teachers and students in their perceptions of quality college teaching and their criteria for judging teaching effectiveness calls for more investigation and verification. It is argued here that many of the mismatches in perceptions can be traced to students' educational upbringing in pre-college education. Therefore, the assumption that quality can be improved in higher education irrespective of what learning styles and habits students bring with them from schools may be unrealistic.

Contrary to the prevailing stance in Oman's higher education, which generally views students' ratings with distrust and suspicion, the present study results appear to provide preliminary support for the use of students' ratings in Oman's universities and colleges as a source of information in teaching evaluation and improvement. It is argued that involving students in the evaluation of teaching is an essential tool in implementing, institutionalising, and enhancing the newly introduced standards in teaching and learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS/ACRONYMS

CTs	Colleges of Technology
CLT	Communicative Language Teaching
GFP	General Foundation Program
HEI	Higher Education Institutions
ICC	Intraclass Correlation Coefficient
IDEA	Instructional Development and Effectiveness Assessment
IELTS	International English Language Testing System
MoHE	Ministry of Higher Education
MTMM	Multi-trait Multi-Method
OAC	Oman Accreditation Council
OQN	Oman Quality Network
ROSQA	Requirement for Oman's System of Quality Assurance
SEEQ	Students' Evaluation of Educational Quality
SET	Student Evaluation of Teaching
SIR	Student Instructional Report
SRF	Student Rating Form
SRT	Student Rating of Teaching
TEFL	Teaching English as a Foreign Language
TESL	Teaching English as a Second Language
TESOL	Teaching English to Speakers of Other Languages
TOEFL	Test of English as a Foreign Language

DECLARATION

I declare that this thesis results entirely from my own work and has not previously been submitted for a degree at this or any other university.

STATEMENT OF COPYRIGHT

ACKNOWLEDGEMENTS

My sincere thanks go to Dr. Sue Beverton and Dr. Tony Harries for their kindness, support, and encouragement. I am grateful to Dr Beverton for her kindness, help and cooperation. I am greatly indebted to Dr. Harries for his patience, advice, and care that helped move the research project through the intermediate and final stages to completion.

Special mention goes to Dr. Robert Coe for his valuable insights and critical support throughout the different stages of the study and for his valuable scholarly comments that greatly shaped my work. I am also grateful to Dr. Patrick Barmby for his kind assistance and advice.

I also wish to recognise the contributions and help of all the Deans, Directors, teachers, and students from the Colleges of Technology in Oman who participated in the study or helped in the data collection. I am grateful to you all.

DEDICATION

To the soul of my father and to my mother.

To my wife and children.

CHAPTER ONE

INTRODUCTION TO THE STUDY

1.0 Introduction

This chapter sets the scene for the study and is divided into a number of sections. After this short introduction, a detailed description of the context of the study, both at the national level and the institutional level, is provided. This is followed by a statement of the problem instigating the present investigation. After that, the purpose of the investigation and its objectives are highlighted and discussed. Then, the research questions guiding the study are stated. The two sections that follow the research questions provide a discussion of the significance of the present study and its scope. This is followed by an explanation of the conceptual framework guiding this investigation. Finally, a section outlining the organisation of the thesis is presented.

1.1 The Context of the Study

This section provides background information about the context of the study. It starts by examining the most recent developments in the higher education system in Oman pertaining to quality assurance and the evaluation of teaching effectiveness in its institutions at the national level. Then, the implications of these developments and reforms on students' role in the evaluation of college teaching are highlighted. Following this, a detailed description of the nature of the program surveyed in this study is presented.

1.1.1 Quality Assurance in Higher Education in Oman

Oman's young higher education system has grown rapidly in the past two decades. Higher education institutions in this Arabian Gulf monarchy are owned and governed by a number of providers, including the Ministry of Higher Education, other governmental entities, and private entities. The programs offered by these institutions are either locally developed or imported, primarily from western countries. The licensing and accreditation systems governing these providers and programs, however, are still in their early stages of evolution and enforcement (Oman Accreditation Council, 2006).

Following the establishment of Oman Accreditation Council in June 2001, the guide *Requirements for Oman's System of Quality Assurance in Higher Education* (locally referred to as ROSQA Document) was issued in 2005. "ROSQA, in effect, is the combination of a number of elements of an overall quality system. It includes a system for classifying institutions of higher education; a qualifications and credit framework; institutional standards; and processes for institutional and program licensing and accreditation" (ibid.: 5).

Under Chapter Four of the ROSQA Document, which is entitled *Standards of Good Practice in Higher Education*, a number of "Quality of Teaching" standards are outlined as requirements for any higher education institution seeking accreditation in Oman. The list is long, therefore only the ones pertaining to teaching effectiveness and/or teacher evaluation are quoted here:

- A comprehensive system for evaluation of teaching effectiveness is in place. Teaching staff develop strategies for improvement of course content and delivery methods. They maintain a portfolio of evidence regarding evaluations, noting strategies for improvement.
- Incentives and rewards are given for outstanding teaching to encourage innovation and creativity, as well as improvement.
- Support and advice are provided for staff to improve teaching through procedures which include induction programmes for new staff, monitoring, supervision and appraisal, and opportunities for professional development.

(Oman Accreditation Council, 2005: 68)

ROSQA goes further and offers a list of "indicators for standards in teaching and learning". The first two of these are:

- Results of survey ratings by students on the relevance and quality of course content, and staff expertise and availability.
- Ratings by students on effectiveness of courses in developing generic competencies defined by the institution.

(Oman Accreditation Council, 2005: 71)

In 2006, ROSQA was revised and developed into a *Plan for an Omani Higher Education Quality Management System* (also referred to as "*The Quality Plan")* in consultation with the higher education sector and international experts. "The purpose of this *Quality Plan* is to give effect to His Majesty's vision by building on current arrangements to establish and maintain an effective quality management system for higher education. It sets out a number of vital goals, objectives and strategies for the Ministry of Higher Education (MoHE), the Oman Accreditation Council (OAC) and the sector at large" (Oman Accreditation Council, 2006: 5). Another major development in quality management and improvement efforts in Oman's higher education was the establishment of the Oman Quality Network (OQN) in 2006.

> The OQN is a collegiate national network of Higher Education Institutions (HEI), supported by the Ministry of Higher Education (MoHE) and the Oman Accreditation Council (OAC). It is concerned with developing a strong and vibrant higher education sector by improving quality in higher education within the Sultanate of Oman. It aims to build a quality-conscious, knowledge-rich higher

> education sector through the sharing of ideas, strategies, research, and practices that inform the pursuit of quality improvement.
>
> (Oman Quality Network, 2008: 3)

The *Quality Plan* has been completed and quality audits have started recently. Because of its enrolment size and perceived role, the General Foundation Program[1] (GFP), which is an entrance program to prepare students for their degree studies, is a current focus of the quality management system (Al Shmeli, 2009: 9). The "GFP standards" were approved in 2008 and accreditation of programs against these standards commenced in the academic year 2009/2010. While the focus and structure of the foundation program may differ from one higher education provider to another, "there is enough in common for Foundation Program standards to be set as a single exercise" (Oman Accreditation Council, 2006: 43). Most GFPs require appropriate levels of English language skills, IT, math, and study skills. Following is a description of the GFP program in the seven Colleges of Technology in Oman surveyed in the present study.

1.1.2 The Nature of the GFP in the Colleges of Technology in Oman

In Oman's general education for pupils aged 6-17, Arabic is used as the medium of instruction. Throughout this stage in public schools, English is taught only as a compulsory foreign language. In tertiary education, however, English is the medium of instruction in most specialisations, including medicine, pharmacy, science, engineering, information technology, and business studies. Only few subjects like history and geography are taught in Arabic.

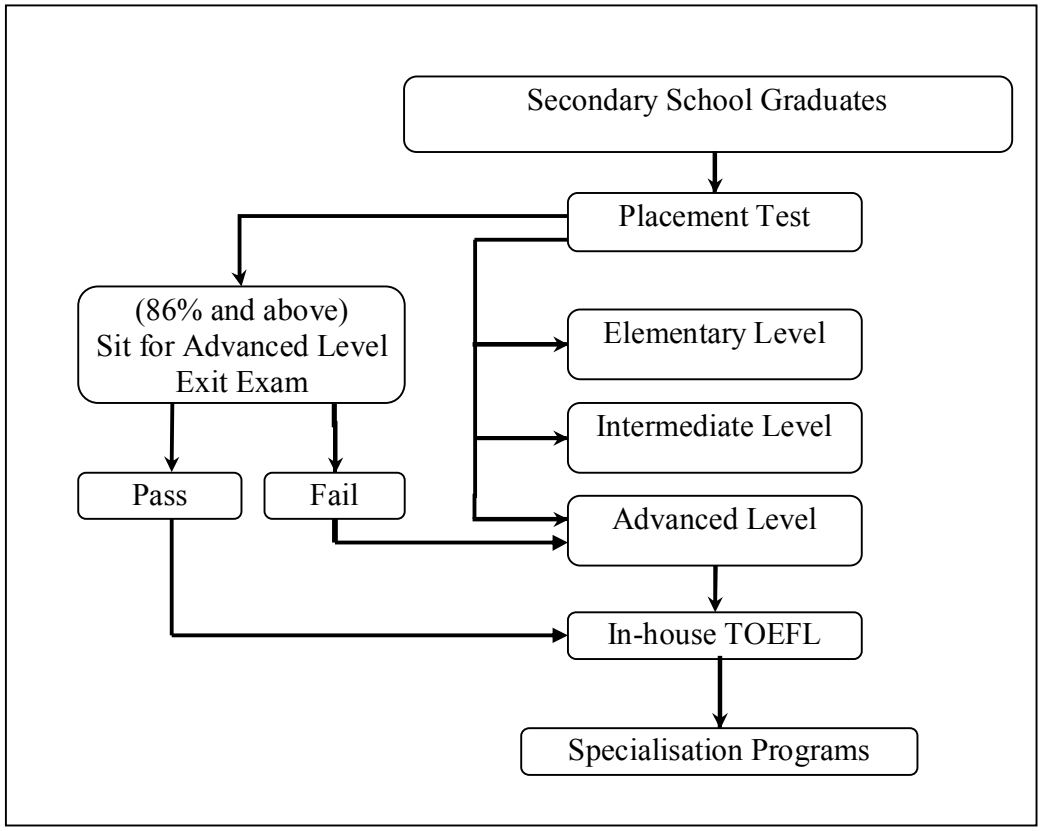
[1] Also referred to as the *Foundation Year* in Oman.

The quality of education students receive in secondary schools in Oman is not yet up to international standards (Issan & Gomma, 2010). Because of limited exposure to English and the quality of materials, instruction, and assessment they receive, students' competency in English upon leaving Grade 12 in public schools is usually far below tertiary education requirements (Al-Issa, 2005; Al-Husseini, 2006). Therefore, upon admission to most higher education institutions in Oman, the majority of students are placed on a GFP. "The GFP is a compulsory entrance qualification for Omani degree programs" requiring students "to achieve English language competency at a level equivalent to IELTS[2] 5.0" (Oman Accreditation Council, 2007: 4). Besides English language, Oman Accreditation Council identifies three other areas of learning outcomes that need to be addressed by the GFPs, namely: mathematics, computing, and general study skills. However, in most GFPs in Oman, English language courses take the bulk of the program.

The structure of the GFP, its organisation, delivery, materials, assessment methods, and progression criteria for students enrolled in it differ from one higher education institution to another. However, in the Colleges of Technology (CTs), where the present study is based, the college contexts and student populations are remarkably similar. The centrally enrolled new students are first required to sit a placement test to determine their entry English proficiency level. Depending on their scores in the placement test, students are placed in one of three levels: Elementary, Intermediate, or Advanced. This is the path taken by the vast majority of students taking the placement test (See Figure 1.1).

[2] International English Language Testing System.

Figure 1.1: Progression of Students through the English Language Program in the GFP in the Colleges of Technology in Oman



Those who do exceptionally well in the placement test (scoring 86% and above), however, qualify to sit the Advanced Level Exit Exam directly. Upon completing this exam successfully, students are entitled to join their specialisation program, provided they meet the all the other admission criteria for that program, including the minimum TOEFL[3] score specified by the prospective department. Those who fail the Advanced Level Exit Exam as part of this qualifying procedure are streamed back into the GFP Advanced Level.

[3] Test of English as a Foreign Language

The three-level English language component of the GFP in CTs is mainly designed to meet students' immediate linguistic and academic needs and prospective labour market requirements (Al-Hinai, 2005). In CTs, the planning and execution of the GFP is primarily the responsibility of the Language Centres.

The researcher worked as a TESOL lecturer in one of the CTs for seven years and then headed a language centre in the same college for 2 years.

1.2 Statement of the Problem

Despite the ROSQA requirements cited earlier, very few higher education institutions in Oman currently use students' ratings (also referred to as students' evaluations of teaching- SETs) as part of their staff evaluation schemes. Every institution seems to have its own reasons for rejecting or limiting the use of students' ratings as a source of data in evaluating teaching effectiveness. From my teaching and management experiences at college level in Oman, the most frequently cited reasons for not using students' ratings have been students' "different and limited" conceptions of quality teaching, "immaturity" and "lack of knowledge and ability" to give reliable and valid judgments about college teaching. Some researchers and educators in the region are also concerned about the applicability and relevancy of some 'imported' rating scales in the Gulf colleges and universities and argue that the educational and cultural upbringing of the students may affect their rating behaviours and prevent them from fully recognising the teaching dimensions underlying these scales.

In institutions where students' evaluations are collected using locally developed rating

forms (see Part C in Appendix 1 for an example rating form locally developed by one of the colleges surveyed in this study), the procedure remains unsystematic, selective and vastly overshadowed by more traditional teacher evaluation schemes, such as classroom visitation and evaluation by the head of department. As evident in Appendix 1, section E, bullet 3, in some institutions, teacher evaluation is mostly confined to new teachers in the probation period, while other members of staff are evaluated "only if there is a serious concern about them". Some universities and colleges even exclude freshmen (also known in Oman as Foundation students) from this practice.

Rating forms are usually developed/adopted without due consideration to the teaching constructs underlying them and without proper validation and testing of their psychometric properties. Locally developed instruments are usually constructed by college administrators or senior teachers and are based on teaching behaviours that are believed to characterise good teaching and effective teachers from the instrument developers' perspective. Nevertheless, no serious attempts seem to have been made to investigate the degree of agreement/disagreement between students' and teachers' perceptions of the characteristics of effective teaching underlying these instruments. Feldman (1988) explicitly cautions of the implications of this, putting the matter strongly, as follows:

> Students' conceptions about good teaching, of course, may or may not match the conceptions of the instructors themselves. … Any … differences in student and faculty views might well contribute to the tensions found in some college classrooms. Moreover, if the faculty and students of a college do not agree as to what constitutes effective teaching, then faculty members may well be leery of students' overall ratings of them, believing their students may use different priorities than they themselves would in arriving at overall evaluations.
>
> (pp. 291-292)

The need for careful consideration of the context and the match/mismatch between students' and teachers' priorities when developing/adopting a rating form may become even greater in multi-cultural environments like the Arab Gulf universities and colleges, where teachers -and students in some countries- come from diverse ethnic, cultural, and educational backgrounds. To exemplify this, research has shown that UAE college students, unlike their multi-national faculty, consider certain human aspects of their teachers, such as respect, flexibility, leniency, and willingness to compromise and inflate grades, as crucial for judging teacher effectiveness (Mercer, 2004; Saafin, 2008).

More importantly, in Oman, as in the rest of the Arabian Gulf countries, higher education institutions are increasingly being modelled on the globally prevailing Western pattern, more specifically the American-style curriculum and program structure. Similarly, quality assurance schemes in these institutions are also being modelled on the Western pattern. Therefore, many of the newly established "American-style" institutions in the Gulf have implemented SET as a means of evaluating the effectiveness of their courses and academic staff (Al-Issa & Sulieman, 2007). As pointed out earlier, however, rating forms are often developed/adopted without consideration to the new context in which they will be used.

> Unfortunately, the American rating scales that are used to conduct these evaluations are often adopted without appropriate modification to take account of the cultural and linguistic backgrounds of the students being assessed.
>
> (Al-Issa & Sulieman, 2007: 302)

Al-Issa & Sulieman (2007), in their examination of students' ratings in the UAE, argue that the cultural and educational upbringing of the students in this part of the world may

affect the way students perceive their ratings of teaching and enhance various possible biasing factors that influence the reliability and utility of students' feedback in teacher evaluation. However, this argument about the transferability of rating instruments across contexts, or lack of it, remains largely at the conceptual level in the Gulf region. No serious attempts seem to have been made to empirically investigate Arab students' ability to identify the teaching dimensions underlying SET instruments originally developed to be used by students in the west.

At a more global level, students from some developing countries were found to view certain dimensions of effective teaching underlying some western SET instruments differently from their western counterparts (e.g. Clarkson, 1984; Lin, Watkins, & Meng, 1995; Watkins & Akande, 1992; Watkins & Regmi, 1992; Watkins & Thomas, 1991). In his investigation of students' perceptions of effective lecturer performance in a Papua New Guinea university, Clarkson (1984) concluded that students from the developing country differ from western students in their perceptions of effective teaching:

> The main difference seems to be that the students from the developing country do not distinguish between the quality of organisation of the lecture and the rapport created between the lecturer and students, whereas students from the developed countries do.
>
> (p. 1386)

While much of the research on college students' ratings of teaching supports the reliability and validity of their judgments of their lecturers' performance, research on freshmen's perceptions of effective teachers and their ability to give valid and reliable ratings of college teaching is very scarce, especially in contexts where English is taught as a second language. Part of the existing literature on students' ratings continues to

cement the stereotypical perception of freshmen as shallow-minded and inexperienced. As Trout (1997) puts it, freshmen do not have the knowledge or the experience to judge the complex dimensions of teaching.

In light of these findings, and the contextual factors pointed out above, a question poses itself about the actual degree of similarity/difference between Omani students and their teachers on their perceptions of the characteristics of effective teaching. The arguments and contextual factors discussed above also invite the question whether college students in Oman, a developing country, are capable of identifying the dimensions of effective teaching underlying a western rating instrument. Bearing in mind the level of scepticism surrounding the ability of Omani freshmen to give reliable and valid ratings of their college teachers mentioned earlier, the reliability of students' evaluations of teaching and the potential effects of various factors hypothesised to bias students' ratings are also areas that require urgent investigation in the Sultanate if students' evaluations of teaching are to be used in making decisions pertaining to quality assurance or educational improvement.

1.3 Purpose of the Study

One can see from the description above that CTs do not use systematic students' ratings as a major source of information in making decisions about staff evaluation and/or staff development for various reasons. Most of these reasons seem to centre on the reliability and validity of student's ratings, and students' perceptions of effective teaching.

However, this attitude towards SET seems incompatible with the newly established quality assurance management system in the country. For example, ROSQA explicitly maintains that high quality teaching standards are important for any higher education institution. Moreover, it implicitly calls for even greater attention to teaching quality in colleges where teaching, as opposed to research, is the prime focus of the institution. According to ROSQA classification criteria, CTs in Oman are categorised as "higher education colleges" emphasising "teaching", and demanding a high level of "teaching effectiveness".

The "Quality Plan' also maintains that the learning outcomes, the students, the academic standards, and the professional development of staff are prime areas of attention in Oman's higher education. Promoting professional development and maintaining high levels of academic standards, however, requires that effective and well-informed schemes for monitoring and improving teaching performance are put in place. In addition, measuring whether the programs are successful in meeting the learning outcomes set for them requires some form of feedback and input from the most important stakeholder in the whole process- students.

To this end, higher education institutions in Oman will find it extremely difficult in the future to justify their decisions to exclude students from this evaluation exercise, whether national quality assurance requirements mandate this involvement or not. What higher education institutions can do, however, is to investigate teachers' fears and doubts about students' ratings and establish the extent to which such fears are grounded on facts or myths. Such research findings, when communicated to the teachers, can help

teachers take a more rational stance towards SET as a source of information on the quality of teaching, with strengths and weaknesses, but no hidden agendas.

This exploratory investigation comes as a contribution to 'break the ice' between educators and SET in Oman by investigating some issues that seem to be a source of concern in students' ratings. As a preliminary investigation that will need to be followed by more research and verification, the present investigation will strive to achieve the following research objectives:

1. Identify the degree of match/mismatch between GFP students' and lecturers' perceptions of effective college teaching.
2. Identify the degree of match/mismatch between GFP students' and lecturers' perceptions of: the factors hypothesised to bias students' ratings; the utility of students' evaluations of teaching; and the role of the student as an evaluator of teaching.
3. Assess the association between students' and teachers' perceptions of effective teaching and various background variables.
4. Identify the dimensions of teaching underlying students' evaluations of teaching in Oman and establish the extent to which these dimensions are similar to or different from those found in western countries.
5. Assess the reliability of students' ratings in Oman and the effect of a number of course, teacher, and student characteristics on these ratings.

1.4 Research Questions

In the light of the description of the context of the problem and purpose of the study given above, this research project will attempt to answer the following main questions:

- **Research Question 1:** To what extent do students' and teachers' perceptions of the importance of various characteristics of effective college teaching match or mismatch?
- **Research Question 2:** To what extent does students' gender have an effect on their perceptions of the importance of various characteristics of effective college teaching?
- **Research Question 3:** To what extent do mediating factors such as teachers' ethnic background and mother tongue have an effect on their perceptions of the importance of various characteristics of effective college teaching?
- **Research question 4:** To what extent do teachers' and students' perceptions of students' evaluations of college teaching match or mismatch on:
    - The effect of the hypothesised biasing factors on students' ratings?
    - The utility of students' evaluations of college teaching?
    - The role of the student as an evaluator of teaching effectiveness?
- **Research question 5:** What dimensions of teaching underlie students' evaluations of teaching in Oman and to what extent are these dimensions similar to or different from the dimensions of teaching identified in the relevant western SET literature?
- **Research question 6:** How reliable are college students' evaluations of teaching in the Omani context?

- **Research question 7:** To what extent do student, lecturer, and course background characteristics influence students' ratings?

1.5 Significance of the Study

Students' ratings of teaching (SRTs)- used interchangeably with students' evaluation of teaching (SET) or students' rating forms (SRFs)- is one of the most common methods of teaching evaluation in higher education (Braskamp & Ory, 1994; Cashin, 1995; Centra, 1993; Marsh, 2007; Seldin, 1999). There is a growing body of research which shows that feedback from students' ratings combined with individual or group consultation is an effective means to improve teaching effectiveness (e.g. Cohen, 1980; Costin, Greenough, & Menges, 1971; Penny, 2004). In America "rare is the … college or university that does not currently use student evaluations of teaching in one way or another" (Centra, 1993: 47). In the UK's higher education "there is an expectation that module or course evaluation is embedded in quality management systems" of the institutions (Rowley, 2003: 142).

Much of the research on SET, however, has been carried out in the USA and Australia. No serious attempts seem to have been made to investigate students' perceptions and use of SET in the cultural and educational context of the Arabian Gulf (Al-Issa & Sulieman, 2007). There appears to have been little attempt also to research the reliability and validity of SET instruments with students for whom English is a second language (ESL) (Pennington & Young, 1989). Furthermore, as hinted at earlier, there seems to be no Gulf-based empirical research which investigates the transferability of rating instruments across contexts- specifically western instruments. These are major gaps in SET research

that this study attempts to fill. Another significant contribution of this study is the translation into Arabic of one of the most widely used western SET instruments, which despite its international reach, does not seem to have been translated into Arabic before.

The present study also draws its significance from the population of students it is targeting and the program under investigation. The sample for the study was exclusively drawn from the GFP. This was purposefully done for the following reasons:

1. There is a wide variety of courses offered in CTs, ranging from academically-focused lectures to work placement. Therefore, approaches in evaluating lecturers' performance may differ significantly from one program to another, which makes comparisons between different programs extremely complicated.
2. GFP students in Oman's higher education institutions constitute a large part of the overall student population. Overall, "It is the single biggest post secondary program in Oman and is taken by 88% of students before their degree studies" (Al Shmeli, 2009: 9).
3. GFPs are and will continue to be important for Oman's higher education. "Even with a robust secondary education system in place, the need for Foundation programs is likely to continue indefinitely for other student cohorts such as international students, mature students or students from lower socioeconomic backgrounds who may not have had access to effective secondary schooling" (Oman Accreditation Council, 2006: 42). Therefore, this study may prove to be beneficial for a big number of higher education providers in Oman and the neighbouring countries.

4. The GFP is considered as a transitional stage between secondary education and higher education. “Students success in higher education is heavily influenced by the effectiveness with which secondary schools prepare them for those higher studies” (Oman Accreditation Council, 2006: 43). For the poorly prepared and for those coming from schools where classes are mostly teacher-centred, for example, GFPs can play a vital role in introducing students to new teaching styles and learning environment, thus preparing them for their specialisation programs. As Al Shmeli (2009) puts it, “[The GFP] helps them make the transition from the traditional learning methods practiced in the schoolroom to the independent mode of study expected in higher education. A student’s future prospects depend on the efficacy and success of the GFP” (p. 9).
5. Freshmen are usually the most vulnerable to discrimination and exclusion from staff appraisal schemes in CTs and are often seen as “immature” and “inexperienced” to judge their teachers’ performance. No studies, however, have been carried out in Oman to back or discredit these perceptions until now.
6. The GFP is well-positioned to play a significant role in preparing freshmen as fair and objective evaluators of college teaching if their ability to evaluate is proven empirically. Involving freshmen in systematic students’ ratings may even help improve students’ attitude toward their studies and clarify their perceptions of the generic characteristics of effective college teaching, which can have a lasting positive effect on their college life and academic performance.
7. Unlike the specialisation programs, the teaching staff in the GFP come from very diverse backgrounds. Students are taught by teachers from all over the world and from varied cultural and educational backgrounds. Perceptions of what

constitutes effective teaching may also be very diverse among teachers and between teachers and students. In the absence of proper research in the subject in Oman, there is a strong need to assess the degree of match/mismatch between different perceptions and the implications of such differences.

8. A large percentage of lecturers in the GFPs are non-native speakers of English. They are allocated the same teaching load and responsibilities as the native speakers of English, who also constitute a significant part of the teaching staff. An emerging body of literature identifies possible biases in students' ratings of teachers whose mother tongue is not English or who are not very proficient in English in contexts where English is the medium or subject of instruction. On the other hand, there are claims that TESOL teachers who can speak the mother tongue of their students are usually given higher ratings. In the Omani context, this area does not seem to have attracted the attention of any studies. More research is needed to determine the existence of the effect of such background variables on teacher rating.
9. Bearing in mind the key role of the GFP as an important pathway into higher education for the majority of Omani students leaving secondary schools, researching the reliability and validity of GFP students' ratings will allow generalisations to be made to other contexts and to higher-level courses beyond the GFP program.

The timing of the study also contributes to its significance. The study comes at a time of major changes and significant reforms in the quality assurance system in higher education in Oman. It is hoped that the findings of this investigation will help inform

some of the policies pertaining to the evaluation of teaching in Oman's universities and colleges.

1.6 The Scope of the Study

With the purpose of the study and the research questions presented above in mind, the scope of the present investigation can be identified in a number of aspects. Firstly, this study does not attempt to offer a single, one-fits-all, conclusive definition of teaching effectiveness, nor does it try to identify a set of characteristics of effective teachers as the only acceptable criteria for judging teaching quality. Moreover, the study does not take as one of its objectives the mission of establishing whether one group's understanding of teaching effectiveness is superior to another, or whether certain teaching methods or techniques are the only ones compatible with effective teaching. Instead, the study is primarily concerned with identifying the degree to which students and lecturers' perceptions of effective teaching and students' evaluation of teaching match or mismatch.

Secondly, the current study also is not concerned with providing a model for staff appraisal or a teaching evaluation system in the institutions sampled in the investigation. Instead, the study aims at establishing whether GFP students can identify the multi-dimensional nature of effective college teaching underlying a SET instrument and the extent to which the factors or constructs of teaching in this instrument are transferable across contexts. In addition, the present investigation examines the inter-rater reliability of students' ratings using the same standardised SET instrument and the extent to which such ratings are influenced by some student, teacher, or course background variables.

Finally, the current study does not aim to promote the use of a specific SET instrument in collecting students' ratings in Oman, nor does it call for the unconditional use of students' ratings as a source of data in evaluating teaching without due care to the contextual factors determining their feasibility and utility in individual institutions. Part of the study, nevertheless, highlights the strengths of standardised SET instruments and the robustness of the procedures and processes used in developing and validating them. Based on the findings and conclusions reached, the study also examines the implications on the evolving quality assurance system in Oman's higher education in general, and the teaching standards and the role of the student as an evaluator of teaching in particular.

1.7 Conceptual Framework

While this study is not explicitly theory-driven, it does draw on various specific concepts and constructs in the domain of higher education. These include teaching and learning in higher education, teacher effectiveness, teacher evaluation and performance management in higher education, the empirical and practical approaches in the design, construction, and evaluation of SET instruments, and quality assurance in higher education. Parts of the study are also set against the broad theoretical knowledge of educational management and policy. The potential links between these concepts and the focal theme of this study- i.e. students' perceptions and evaluation of the effectiveness of college teaching - can be complex and multi-directional. In addition, the existence and strength of these links in a particular context may be determined by the strength of the relationship between research and policy or practice in that context. To illustrate this, Teddlie, Stringfield and Burdett (2003) offered a "rational model" which illustrates potential links between teacher and school effectiveness research, teacher evaluation,

staff development, and teacher and school improvement. According to the theoretical links in their model, Teddlie *et al.* argue that teacher evaluation can inform staff development and generate teacher improvement, which is expected to bring about positive changes in student learning and lead to school improvement.

While arguing for the theoretical viability of these links, Teddlie *et al.* (2003) admit that their existence in reality varies widely by context, with some educational systems having more 'loose links' than others. They attribute the loose links among the different constructs highlighted above to the disconnect between educational research and policy and point out a number of factors that lead to this situation. Among these factors are: (1) the separation of research and practice in educational improvement; (2) the "intellectual and/or academic traditions" in some countries which consider teaching to be a "craft" that cannot be investigated scientifically; and (3) school autonomy and teacher union resistance to meaningful staff appraisal in some contexts.

However, even in contexts where the link between effectiveness research and staff development is strong, another "criticism levelled at teacher development activities is that they are not, for the most part, theory-based" (Penny, 2004: 9). Penny adds that such lack of theoretical base can be attributed to the fact that such activities are usually more concerned with teaching practice and demonstrating 'what is useful' and what works rather than designing theoretically sound teacher development programs. Such lack of theoretical base can also be attributed to the exclusive emphasis on theory verification rather than theory generation dominating research in higher education (Conrad, 1982).

When considering students' evaluation of their teachers, most SET instruments are in practice designed on a mixture of logical, pragmatic, and occasionally psychometric considerations (Marsh & Dunkin, 1997). In identifying, constructing, and evaluating the multiple dimensions of effective teaching in SET instruments in particular, Marsh and Dunkin identified three interlacing approaches. The first approach is empirical. This includes factor analysis and multitrait-multimethod (MTMM) analysis. The second approach is logical, which mainly analyses the characteristics of effective teaching as perceived by instructors and students supported by reviews of the research in the field. The third approach is based on the theory of teaching and learning. "In Practice, most instruments are based on either of the first two approaches- particularly the second" (Marsh, 2007: 322). For example, *Students' Evaluation of Educational Quality* (SEEQ), a widely used standardised American SET instrument, while having well-developed psychometric properties and well-defined factor structure based on multiple factor analyses and logical investigations of the characteristics of effective teaching from teachers' and students' perspectives, measures constructs which are less well rooted in student learning/teaching theory (Coffey & Gibbs, 2001).

In the field of teaching effectiveness in higher education, a factor that adds to this disconnect between research, policy, and practice is the lack of agreed upon definition of teaching effectiveness. Marsh (2007) stresses the importance of integrating theory, research, and practice in measuring the components of teaching effectiveness as a starting point towards designing valid measurements and useful SET programs. However, he admits that teaching effectiveness is a "hypothetical construct" which is difficult to measure:

> An important, unresolved controversy is whether the SET instruments measure effective teaching or merely behaviors or teaching styles that are typically correlated with effective teaching. … Nevertheless, there is little or no systematic evidence to indicate that any of the typical SET factors is negatively related to measures of effective teachings. … Because teaching effectiveness is a hypothetical construct, there is no measure (SETs or any other indicators) that IS effective teaching- only measures that are consistently correlated with a variety of indicators of teaching effectiveness.
>
> (pp.322-323)

To this end, the lack of a unified definition of effective teaching does not justify disregarding the huge body of research on the utility of various indicators of teaching effectiveness, including SET, in educational improvement efforts. Such arguments, however, do not appear to inform policy and practice in many higher education institutions. In the context of the study, personal conceptions and collective wisdom about what works, what is useful, what is effective, and who can evaluate all of this seem to speak louder than research evidence in board meetings and appraisal interviews. As illustrated in Figure 1.2, in a context of change, this separation between educational research and practice/policy can hinder the implementation and institutionalisation of educational change and reforms. As suggested by Fullan (2001), planning and managing educational change is a three-phase process which involves initiation, implementation, and institutionalisation.

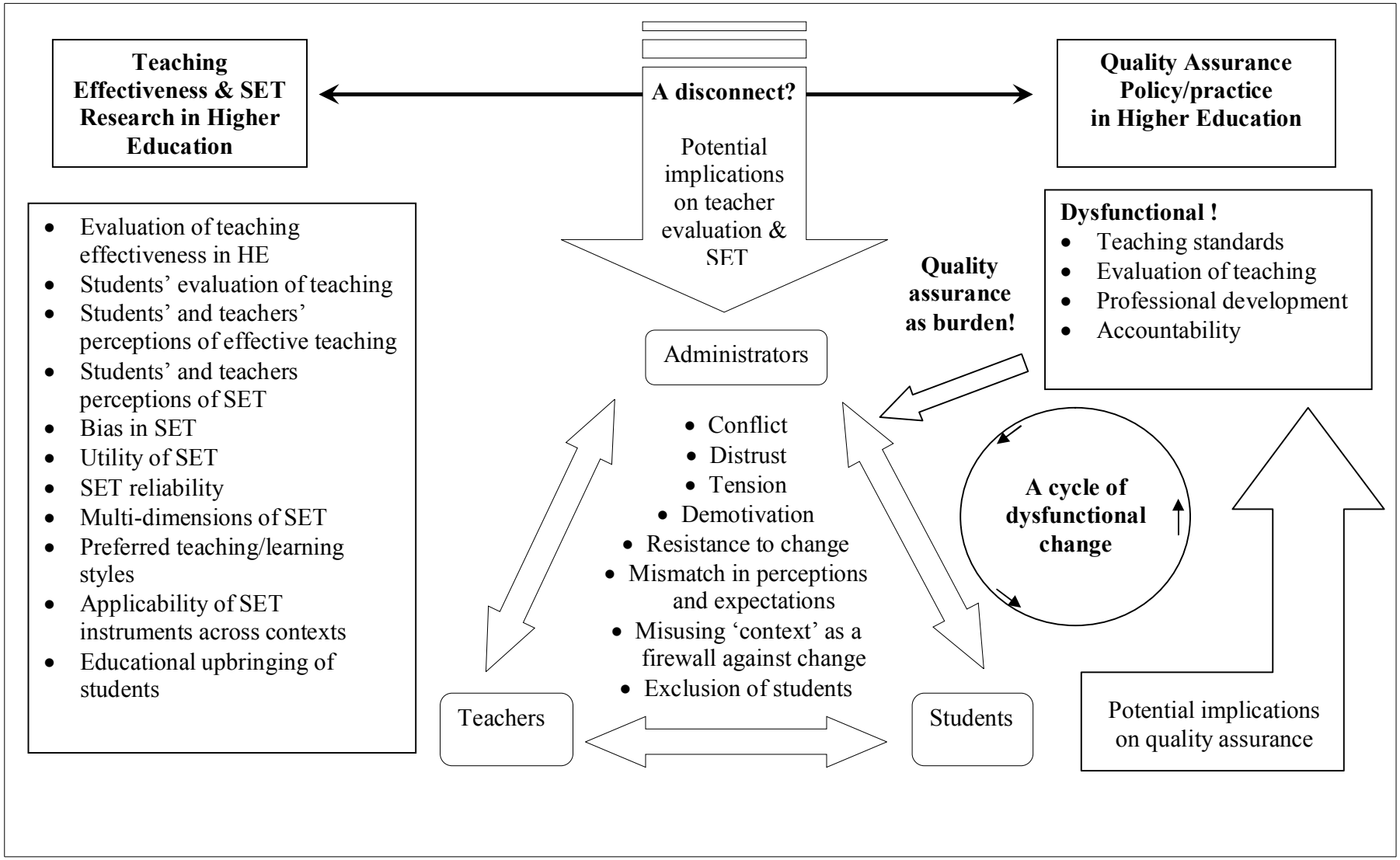
In the context of Oman higher education, great hopes are placed on the newly initiated quality management system as a whole, and the academic standards and teaching quality indicators embedded in it in particular. The new emphasis on the student and the learning outcomes as opposed to the teacher and traditional knowledge transmission approaches are hailed as two of the most significant prospective improvements.

However, this new emphasis on the student does not seem to be in proportion with student's role in the evaluation of teaching/learning, both existing and prospective. In the 'Quality Plan' mentioned earlier, there is little or no mention of how students will be involved in evaluating their learning experiences and in providing feedback to their institutions on the quality of teaching they receive. This seems to be in contradiction with the new philosophy embedded in the new quality assurance system which considers the learner as the focus of the educational process.

Furthermore, the initiation of standards in Oman's higher education does not necessarily mean that the competencies required to implement these standards are readily available (Carroll, Razvi, & Goodliffe, 2009). With specific reference to the GFP in Oman, Carroll *et al.* (2009) highlight the calibre of the teachers as a critical factor in the success of the implementation stage:

> Whereas the development of standards involves the best available expertise in each subject, the implementation of standards involves all academicians involved in the teaching of GFPs. Working with explicit academic standards for student learning outcomes requires a heightened praxis by teachers, because turning standards into practice is complex. …
>
> These challenges, in turn, shine a light on the calibre of academicians and the adequacy of professional development opportunities for those staff. This is a particularly important issue in a country which relies heavily on foreign academicians employed on fixed term contracts. …They are expected to arrive in the country work-ready and, as such, do not necessarily have access to professional development opportunities. …This will limit the capability of HEIs to prepare for accreditation of their programs.
>
> (pp.6-7)

Figure 1.2: Conceptual Framework



The institutionalisation of academic standards and, consequently, of professional development programs, however, require systematic and constant input from the students as key stakeholders in the educational process. Here it is theorised that alienating students from the evaluation of teaching and disregarding their role in enhancing academic standards may seriously affect the implementation and functioning of teacher professional development programs and teaching quality assurance in higher education institutions. Teacher evaluation policies and practices that are based on mere preconceptions and speculations about what students can or cannot do and which do not take account of the empirical evidence in the field may give rise to distrust and conflict in the college classroom. Understanding how our students view effective teaching, what influences their perceptions of our teaching, and their preferred teaching/learning styles, on the other hand, can give us valuable insight in designing good and relevant professional development programs for the teachers and in successfully implementing quality assurance measures.

1.8 Organisation of the Thesis

The thesis consists of nine chapters. After this introductory chapter, a review of the relevant literature is presented in three chapters and in a funnelling-effect pattern, from general to specific. Chapter 2 opens the review by examining the evaluation of teaching effectiveness in higher education. This is followed by a discussion of the research in the matches and mismatches between students and teachers' perceptions of teaching effectiveness in Chapter 3. Chapter 4 closes this review with a discussion of the research on students' ratings of college teaching and the reliability, validity, and utility of

students' evaluations of teaching and the applicability of standardised rating instruments across contexts.

Chapter 5 is the research methodology and design chapter. This chapter presents a description and discussion of the research methods and design used in the two stages of the investigation: the qualitative email-based exploratory study, and the primarily quantitative main study. For both stages, the chapter offers a detailed account of the research methods and design employed in collecting and analysing the data. This chapter also presents a description of the research ethics guiding the present study.

Following this are three chapters for the study data presentation and discussions. Chapter 6 presents and discusses the findings on the matches and mismatches between students and teachers' perceptions of the importance of various characteristics of effective teaching. Chapter 7 presents and discusses the results of the analysis of the similarities and differences between students and faculty's perceptions of SET ratings: their hypothesised biasing factors, their utility, and the role of the student in the evaluation of teaching. The data presentation and discussion section of the thesis closes with Chapter 8 which presents and discusses the factor analysis and inter-rater reliability analysis of SEEQ in Oman and the potential influences of certain student, lecturer, and course background characteristics on students' ratings.

Finally, the thesis concludes with Chapter 9 which sums up the findings in the results section and discusses the implications of these findings on the policy and practice in the evaluation of teaching effectiveness and quality assurance and control of teaching

standards in higher education in Oman. After Chapter 9, the appendices and references are presented.

CHAPTER TWO

TEACHING EFFECTIVENESS IN HIGHER EDUCATION

2.0 Introduction

This chapter explores teaching effectiveness in higher education in a number of steps. Firstly, the concept of teaching effectiveness and the characteristics of effective teachers in higher education will be discussed. Secondly, classroom teaching effectiveness as a key criterion in overall faculty evaluation in higher education, and its position in relation to other criteria, will be examined. Then, the different sources of information used in the evaluation of teaching effectiveness in higher education, which are in actual use and have been cited repeatedly by the relevant literature, will be investigated. Finally, the role of the students as evaluators of teaching and the effect of their perceptions of effective teaching and their conceptions of the utility of students' evaluations on the rating exercise will be addressed.

2.1 The Concept of Effective Teaching: Sources of Variability

Researchers and educators are divided on the issue of defining effective teaching. Various factors contribute to this division. Some factors are inherent in the multidimensionality, complexity, and variability of teaching itself. Other factors are related to the different theories underlying learning/teaching which guide some of the research on teaching effectiveness. A third group of factors are related to the conceptions of teaching work and how researchers as well as professionals view the nature of teaching.

2.1.1 The Multidimensionality, Complexity, and Variability of Teaching

Some researchers (e.g. Adams, 1997; Brown, 1996; Marsh & Dunkin, 1992; North, 1999; Patrick & Smart, 1998) stress that teaching is multidimensional and complex, and therefore, it is difficult to construct a one-fits-all definition of effective teaching. Others (e.g. Centra, 1993; Hativa, 2000) also argue that student learning at college level is a far more complex activity and can be affected by several factors besides teaching effectiveness, such as student efforts, aptitudes, learning styles, and preferences for teaching styles. In this light, teaching may be regarded as an activity that is difficult to measure systematically in a pattern that allows comparisons between individual teachers, partly because definitions of teaching effectiveness may depend on the individual's explicit or implicit theory of how students learn.

Besides the difficulties caused by the multidimensionality and complexity of teaching in defining teaching effectiveness, there is also the assumption that teaching tasks can be quite personal and idiosyncratic at times, reflecting variety in the teaching/learning setting and the styles, needs, and learning preferences of students. " Good teachers are good for different reasons…What makes one teacher good (an effective taskmaster) may not be true of the next one (an inspirer) or still another (a subject matter authority)" (Peterson, 1995: 6-7). Teaching and learning are also dynamic activities that take place in constantly changing environments and, consequently, sometimes require radical changes in our approach to teaching and learning. It is an activity that is usually ruled by "the law of unintended consequences" as Elton (1996) calls it.

2.1.2 Theories of Learning/Teaching

Fuhrmann and Grasha (1983), identify three definitions of effective teaching that are based on three different learning theories:

1. The *behaviourist approach*: This is a very systematic theory of teaching that is prevalent in college teaching and is mostly instructor-centred and lecture-based. According to this approach, "effective teaching is demonstrated when the instructor can write objectives relevant to the course content, specify classroom procedures and student behaviours needed to teach and learn such objectives, and show that students have achieved the objectives after exposure to the instruction" (ibid.: 287).
2. The *cognitive theory-based approach*: This theory emphasizes problem solving and critical thinking skills. According to this approach, "effective teaching is demonstrated when instructors use classroom procedures that are compatible with a student's cognitive characteristics, can organize and present information to promote problem solving and original thinking on issues, and can show that students are able to become more productive thinkers and problem-solvers" (ibid.: 287-288).
3. The *humanistic theory-based approach*: This approach promotes self-initiated leaning where teachers take the role of model learners rather than the role of expert. "Humanistic teaching is effective when teachers can demonstrate that students have acquired content that is relevant to their goals and needs, that they can appreciate and understand the thoughts and feelings of others better, and that they are able to recognize their feelings about the content" (ibid.: 288).

Centra (1993) concludes that although there are other theories that offer other assumptions how teachers should teach and how students learn, the three theories mentioned above probably sum up most viewpoints.

2.1.3 Conceptions of Teaching Work

Related to the arguments above about the different theories underlying teaching and learning and their perspective on effective teaching is the argument about the different conceptions of teaching work and how these conceptions affect teaching evaluation. According to many authors (e.g. Mitchell & Kerchner, 1983), there are four main conceptions of teaching work: labour, craft, profession, or art. It is argued that the evaluation of teaching varies depending on one's conception of teaching (Darling-Hammond *et al.*, 1983). Although *pure* prototypes of teaching work do not exist in reality, every teacher evaluation technique is implicitly based on assumptions about teaching work and the role of the teacher in the administrative structure of the educational organisation (ibid).

When conceived as *labour*, ideal teaching tasks are claimed as "rationally planned, programmatically organized, and routinised in the form of standard operating procedures" by programme administrators (Mitchell & Kerchner, 1983: 35). Teacher evaluation under this conception involves direct inspection of the teacher's work by the school administrator who is seen as the teacher's *supervisor*.

Under the conception of teaching as *craft*, good teaching is perceived as "requiring a repertoire of specialized techniques. Knowledge of these techniques also includes

knowledge of generalized rules for their application" (Darling-Hammond *et al*, 1983: 291). Here also the teacher works under the close supervision of the administrator, who is seen as a *manager*.

When viewed as *profession*, "[effective] teaching is seen as not only requiring a repertoire of specialised techniques but also as requiring the exercise of judgement about when those techniques should be applied" as informed by a body of theoretical knowledge (Ibid: 291). Under this conception, the head of school is seen as an *administrator* whose job is to make sure that all the necessary resources are made available to the teachers to carry out their work.

Under the conception of teaching as *art*, teaching tasks and techniques are rather personalised than standardised. Intuition, creativity, improvisation, and the personal resources, skills, and insights of the teacher become important factors in teaching effectives. Here the school administrator acts as a *leader* whose responsibility is to encourage the teacher's efforts.

As mentioned earlier, the four conceptions explained above are ideal types that do not necessarily exist in pure forms. However, these conceptions of teaching work embody different definitions of 'good' and successful teaching which imply different ways by which information about performance is collected and judgments about teaching effectiveness are made.

2.2 Towards a Working Definition of Effective Teaching

Despite all these difficulties facing researchers in reaching a precise and agreed-upon definition of effective teaching, some researchers are of the view that although teaching effectiveness is difficult to define and measure, it should not be totally ignored simply because we do not have a neat definition for it. Some researchers (like Abrami, d'Apollonia, & Rosenfield, 1997; Ellett & Teddlie, 2003; Hativa, 2000; ; McKeachie, 1997a; Ramsden, 1992, 2003; Seldin, 1998, 1999) focused more on student learning as an important indicator of good teaching and tried to define effective teaching from this perspective. Hativa (2000), for example, defines effective teaching as "teaching that brings about effective and successful student learning that is deep and meaningful" (p.11). Targeting more precise aspects of student learning, Abrami *et al.* (1997) give a more focused definition which emphasises students' cognitive skills, attitudes and interests in the subject matter, and interpersonal skills and stress that effective teaching entails "the positive changes produced in students in relevant academic domains including the cognitive, affective, and occasionally the psychomotor ones" (p.324).

Other researchers, challenged by the hypothetical construct of teaching and its complexities, attempt to describe the teaching skills and learning environment that effective teaching creates or promotes. Penny (2004) concludes that effective teaching is teaching which creates a learning environment which promotes discovery, knowledge construction, creativity, critical thinking, and life-long learning skills among students. "It is not teaching that merely transmits knowledge causing students to passively absorb information" (Ibid: 16). Arreola (1986, 1989), on the other hand, argues that for a

definition of good college teaching to be complete, it should include three main dimensions: content expertise, instructional delivery skills and characteristics, and instructional design skills. Highlighting the importance of both the product of effective teaching (i.e. student learning) and the procedures (i.e. what teachers do), Centra, Froh, Gray, & Lambert (1987), define effective teaching as that which "produces beneficial and purposeful student learning through the use of appropriate procedures" (p.5).

Despite the controversy over the concept of teaching effectiveness evident in the discussion above, a reasonable degree of agreement exists about the generic characteristics of effective teachers. Several major reviews of the research on the characteristics of effective teaching (e.g. Feldman, 1988; Ramsden, 2003; Schaeffer, Epting, Zinn, & Buskit, 2003) revealed that a number of characteristics and teaching behaviours are consistently associated with effective teachers. These are discussed in the following section.

2.3 The Characteristics of Effective Teachers

There is a huge body of literature which presents consistent findings on what constitutes effective teaching. While this literature points to more agreements than disagreements among researchers on the generic characteristics of effective teaching, understandably, no single list of the characteristics or qualities of effective teachers, however, has yet been developed to everyone's satisfaction across contexts and courses (Penny, 2003, 2004; Seldin, 1999).

Some studies exclusively used college students' perceptions as their data source. One important assumption that underlies such studies is the belief that we cannot improve teaching unless we are aware of what our students think of it (Richardson, 1998). From as early as 1950s, research findings have shown that students' ranking of the qualities of a good instructor correlated positively with the rankings of alumni (Costin, Greenough, & Menges, 1971). Crumbley, Henry, & Kratchman (2001) specifically researched undergraduate and graduate students' (n= 530) perceptions of effective teaching and found that students identified the following teacher qualities as factors that are likely to improve students' ratings of their college teachers: teaching style (88.8%), presentation skills (89.4%), enthusiasm (82.2%), preparation and organization (87.3%), and fairness in grading (89.8%). In a more recent study, Okpala and Ellis (2005) investigated the perceptions of U.S college students of teacher quality components. The following five key characteristics emerged: caring for students and their progress, instructional skills, knowledge of subject area, dedication to teaching, and communication skills.

Other studies examined data obtained from both faculty and students. Feldman's (1988) review of thirty-one studies in which students and faculty members were asked to specify the characteristics they believed to be important to good teaching and good teachers revealed extensive similarities between the findings of these studies. Students and faculty members identified the following qualities:

- Sensitivity to class level and progress
- Preparation and organization of the course
- Mastery of subject area
- Enthusiasm for the course and for teaching
- Clarity and expressiveness
- Availability and support for students

- Fairness and impartiality in evaluation and grading
- Quality of evaluation and examinations

In more recent studies and reviews, other researchers (e.g. Ramsden, 2003; Raymond, 2008; Schaeffer *et al.*, 2003) echoed almost similar findings, reporting strong similarities between teachers' and students' perspectives about effective teaching. "When students are asked to identify the important characteristics of a good lecturer, they identify the same ones that lecturers themselves do: organization, stimulation of interest, understandable explanations, empathy with students' needs, feedback on work, clear goals, encouraging independent thought" (Ramsden, 2003:87). A more detailed discussion of the match/ mismatch between college students' and faculty's perceptions of the characteristics of effective teaching is presented in Chapter 3.

A third type of studies attempted to investigate the qualities of effective college teachers as perceived by heads of university departments. Kane, Sandretto, & Heath (2004) asked heads of university science departments to nominate lecturers whom they considered excellent teachers. Five broad categories of characteristics of "excellence" resulted from this study. These were: knowledge of the subject, pedagogical skill, interpersonal relationship, research/teaching nexus, and personality.

Researchers like Centra (1993), d'Apollonia & Abrami (1997), Kolitch & Dean (1999), and Penny (2003), however, argue that many of the characteristics of effective teaching identified by Feldman's (1988) review and others are more representative of didactic teaching, not reflecting the diversity in teaching styles, disciplines, and contexts in

college classrooms. As Centra (1993) puts it, these characteristics "presuppose a lecture or lecture-discussion method. [Participants'] responses, therefore, really address the characteristics of effective didactic teaching" (p. 41) and overlook other alternative methods. d'Apollonia & Abrami (1997) agree with this view and add:

> These alternative methods range from interactive seminars, laboratory sessions, and cooperative learning to more independent methods such as computer-assisted instruction, individualised instruction, and internships. Whether the definitions of instructional effectiveness are based on instructional products, processes, or the causal relationships between the two, they do not necessarily generalise across these other instructional contexts.
>
> (p. 3)

Centra (1993) contests that some good teaching qualities are more easily measured than others and, consequently, may be overemphasised or given more weight than their actual effect. He also criticises the methodological approach of some of these studies and points out that some studies failed to recognise the possibility that effective teaching may be affected by factors such as teacher's individual style, academic discipline, academic level, and even individual student. For example, effective teaching at the graduate level may not necessarily be the same as effective teaching at the undergraduate level and teachers who do well at one level may not do well at the other. To help accommodate such diversities in the teaching/learning context, Centra (1993) and other researchers (e.g Braskamp & Ory, 1994; Penny, 2003) proposed the development of a bank of items that characterise effective teaching, reflecting a wide variety of teaching/learning methods, disciplines, and contexts, which then can be added to a set of core items in teacher evaluation instruments, such as student rating forms. The system is sometimes referred to as the *Purdue's Cafeteria System* (Centra, 1993). Such arrangements can help alleviate some of the drawbacks associated with standardised rating instruments. However, the questions remain, though, where does actual classroom

teaching, as an index of faculty performance, stand in comparison to other performance indicators in overall faculty evaluation as evident in the relevant research literature? And what sources of information are used in evaluating the overall performance of faculty?

2.4 Teaching Effectiveness as an Important Factor in Overall Faculty Evaluation in Higher Education

In considering a college professor or teacher for tenure, promotion, or retention, higher education institutions today examine a wide range of factors. Classroom teaching effectiveness, however, remains the most important indicator of faculty performance (Seldin, 1999). In three nationwide surveys conducted in 1978, 1988, and 1998 in the United States to examine the wide range of practices in faculty performance evaluation in liberal arts colleges, Seldin found that "almost to a person, the academic deans chose classroom teaching as the most important index of faculty performance" (Seldin, 1999: 5). The survey listed 13 factors used in faculty overall evaluation and asked the deans to label each one as "major factor", "minor factor", "not a factor", or "not applicable". Although given lower rankings, other factors, such as student advising, committee work, research, and publication were also among the top rated. Table 2.1, adapted from Seldin (1999), summarises the most important findings for the top seven ranking factors in the three surveys.

Table 2.1: Frequency of Use of Factors Considered in Evaluating Overall Performance in Liberal Arts Colleges in the U.S, 1978, 1988, and 1998

Factors	**Frequency of use as a 'major factor' in evaluating overall faculty performance**		
	1978	**1988**	**1998**
Classroom teaching	98.8%	99.8%	97.5%
Student advising	66.7%	64.4%	64.2%
Campus committee work	48.8%	54.1%	58.5%

Length of service	49.9%	43.9%	43.8%
Research	24.5%	38.8%	40.5%
Publication	19.0%	29.4%	30.6%
Personal attributes	38.4%	29.4%	28.4%

Seldin (1999) concludes that things did not change very much with regard to the evaluation of overall faculty performance over the 20 year period between 1978 and 1998, despite the rising calls among policy makers and institutional governing boards for greater accountability in higher education. He also stresses that the importance of classroom teaching as a leading factor, followed by student advising, campus committee work, and length of service in descending order by 1998 scores, continued to be strong throughout the period of the study.

The degree of emphasis on teaching effectiveness in faculty performance evaluation, however, "is a function of the mission of the institution" (Morreale, 1999: 116). According to Morreale, higher education institutions can be divided into three categories: research universities, liberal arts colleges and community colleges, and comprehensive universities, with more emphasis placed on the scholarship of teaching in the second type. Morreale's classification raises the question whether Seldin's findings about teaching effectiveness as a key factor in overall faculty performance evaluation in liberal arts colleges can be generalised over the contexts of research universities and comprehensive universities. Regardless of the weight assigned to teaching in each type of institutions, though, there is even a more important question to ask: what information sources do deans and program administrators use to evaluate teaching competence?

2.5 Sources of Information Commonly Used in the Evaluation of Teaching Effectiveness

Darling-Hammond *et al.* (1983) define the evaluation of teaching as gathering, interpreting, and using data to judge the worth of teaching. A number of sources of information have traditionally been used for this purpose. Centra (1977), cited in Centra (1993), reports the findings of a survey he conducted in 1976 where 453 department heads were asked to rank the current use and importance of fifteen types of data sources in evaluating teaching (shown on Table 2.2). Centra's list of possible sources of data in teacher evaluation is of great importance, as it "summarises most of the kinds of data used in both previous and subsequent research" (Cashin, 1989: 1).

Almost the same list was used later by Seldin in three similar surveys reported in Seldin (1999). To find out how deans assessed teaching effectiveness and what information sources they used, Seldin's nationwide surveys of 1978, 1988, and 1998 asked the deans to indicate how frequently they used 15 possible sources of information (shown in Table 2.2) to assess faculty teaching performance. The findings of Centra's and Seldin's surveys are summarised in Table 2.2 overleaf.

Table 2.2: A Summary of Four Survey Findings on Information Sources Used in Evaluating Teaching Effectiveness

Information Source *	Centra 1976	Seldin 1978	Seldin 1988	Seldin 1998
	Ranked by "current use & importance"	*Frequency of use "Always Used"*		
Systematic student ratings	2nd(tied)	54.8%	80.3%	88.1%
Evaluation by department chairs	1st	80.3%	80.9%	70.4%
Evaluation by dean	6th	76.9%	72.6%	64.9%
Self-evaluation or report	9th	36.6%	49.3%	58.7%
Committee evaluation	4th	46.6%	49.3%	46.0%
Colleagues' opinions	2nd (tied)	42.7%	44.3%	44.0%
Classroom visits	12th	14.3%	27.4%	40.3%
Course syllabi and exams	7th	13.9%	29.0%	38.6%
Scholarly research/publication	**	19.9%	29.0%	26.9%
Informal student opinions	5th	15.2%	11.3%	15.9%
Alumni opinions	13th	3.4%	3.0%	9.0%
Grade distribution	**	2.1%	4.2%	6.7%
Long-term follow-up of students	14th	2.2%	3.2%	6.0%
Student examination performance	11th	2.7%	3.6%	5.0%
Enrolment in elective courses	8th	2.7%	1.2%	1.5%

** In descending order by Seldin's scores of 1998 survey*
*** These two sources did not appear on Centra's survey of 1976. Instead, two other sources were used, namely: "Teaching improvement activities" (10th), and "Videotape of classroom teaching" (15th).*

It is evident from the data above that there were significant changes during the period between 1978 and 1998 in the methods employed by the surveyed institutions in evaluating teaching effectiveness. Four sources of data, namely: systematic student ratings, self-evaluation, classroom visits, and course syllabi and exams, changed by 20 % or more in that period and became more widely used (Seldin, 1999). It is also clear from the table above that ratings by students, the department chair, and the dean remained predominant throughout the period of the study, with student ratings becoming the most widely used source of information on teaching effectiveness in 1998. As far as

the use of students' ratings is concerned, this finding confirms the results of Seldin's (1993a) earlier research. In another study which tracked the use of students' evaluations of teaching in 600 American colleges between 1973 and 1993, Seldin (1993a) found that the use of students' ratings in evaluating teaching effectiveness increased from 29 percent to 86 percent.

It remains unclear, nevertheless, whether these institutions use this data for developmental purposes, evaluative purposes, or a combination of both. It also remains unclear how deans, department chairs, colleagues, and committees can make sound judgments about a professor's effectiveness as a teacher in the absence of classroom visits or student ratings. Cashin (1989) expresses his reservations regarding these four sources of data and argues that "[these] are not data but evaluators" (p.1) who obtain their data from other sources on the list, such as classroom visits and systematic student ratings.

In his expanded definition of college teaching, Cashin (1989) identifies several possible sources of data in evaluating college teaching in a more detailed fashion than Centra's and Seldin's surveys suggested (see Appendix 2). He proposes a grid in which seven *areas of teaching* are listed along the vertical dimension. These are: subject matter mastery, curriculum development, course design, delivery of instruction, assessment of instruction, availability to students, and administrative requirements. Under each area are listed some specific aspects or facets of teaching. Along the horizontal dimension are listed eight sources of data in evaluating college teaching- most of which are people working in the educational organisation who can supply or evaluate data. These sources

are: self, files, students, peers, colleagues, chair/dean, administrator, consultant, and others. In the intersecting cells are specific examples of possible types of data available to teaching assessors.

Cashin lists the instructor himself or herself first "because he or she has knowledge about all of the areas of teaching, and there are some things that only the instructor may know" (Cashin, 1989: 2). According to Cashin, the instructor's self-report should primarily be descriptive rather than evaluative. Student interviews, comments, and letters are also considered important by Cashin as sources of data about a teacher's performance. However, he does not recommend them as routine methods to collect student feedback, partly because they are time consuming, but mainly because different students' views may be poorly represented, as data typically comes from volunteers. Student ratings, on the other hand, are considered by Cashin as the only primary data that are systematically collected at many higher education institutions. A detailed discussion of student evaluations of teachers in higher education is presented in Chapter 4.

It is also clear from Cashin's classification that Cashin draws a distinction between "peers" and "colleagues". He restricts the term "peers" to "faculty members knowledgeable in the subject matter", while he uses the term "colleagues" to refer to " all of those faculty … familiar with higher education's academic enterprise but not knowledgeable of the specific subject matter" (Cashin, 1989: 2-3). Judgments of peers and colleagues based on classroom observation, for instance, can be a valuable source of information in various teaching areas. As with peers and colleagues, deans and

department heads, Cashin stresses, can base their judgment on data sources such as classroom observations and visits. Although their use in personnel decisions remains controversial, "classroom visits have won increased popularity as an evaluative tool" (Seldin, 1999: 19), especially when conducted by trained observers. In addition, deans and department heads can also make use of department/college files which usually contain information relevant to the teacher's fulfilment of administrative requirements, as well as information about his or her contributions to the instructional programme.

Another important source of information about teaching effectiveness highlighted in Cashin's grid is the instructional consultant, who is usually a faculty member or other trained professional whose job is to help teachers improve their teaching practices and techniques. Several researchers (e.g. Marsh & Roche, 1997; Kwan, 2000; Penny, 2003, 2004) have argued that feedback given to teachers on their teaching performance- such as students' ratings- coupled with consultation results in substantial improvement in teaching. Cashin (1989), however, cautions against the conflict of interest that may arise if this consultant is asked to make a judgment for personnel decisions and strongly recommends that consultants be asked to offer judgments for improvement only.

2.6 Students as Evaluators of Teaching Effectiveness

In keeping with the original purpose of this study explained in Chapter One, and also because of the growing importance of student ratings of teachers as the main source of teaching evaluation data as evident in the studies discussed above, this closing part of this Chapter will focus on the role of students in the evaluation of teaching effectiveness. Unlike the instruments used in the process of student evaluations of teaching (SET) and

their psychometric properties, the role of students in this process has been largely ignored by research and very few attempts have been made to investigate this issue (McKeachie, 1997b; Feldman, 1998; Kwan, 2000). Kwan (2000) argues that:

> Any use of student evaluations [of teaching effectiveness] must be based on the assumption that students are willing and able to provide valid judgements about the teaching they have received … .
>
> [Therefore,] it is of vital importance for us to know more about the role of the students in the process of evaluating their teachers or courses. For Example, we need to understand more about how students make sense of evaluations, what influences their attitudes towards evaluations, how they react to the rating process, and what goes on in their minds while making their ratings, etc.
>
> (pp. 20-21)

Kwan's (2000: 118-124) qualitative investigation of the attitudes and rating behaviours of university students in teaching evaluations reveals a number of important findings. These are summarised below:

- Though students demand to have the chance to evaluate their teachers, they generally believe that their ratings have little impact on the improvement of teaching practices.
- Students demonstrated different understandings for the purpose of teacher evaluation and identified different motives for their participation in the rating process.
- Students' understanding of the purpose of the process of rating their teacher's performance and the role they play in this process may have a strong effect on how seriously they respond to the task of rating.
- Students' attitudes about the rating process are strongly affected by the context of the evaluations. Long forms, irrelevant or ambiguous questions,

and unclear aims of evaluation are among the factors which may influence students' attitudes towards the evaluation exercise.

- Students use a variety of strategies to decide on what ratings to give to their teachers. Their ratings can be criterion-referenced (i.e. the extent to which the teacher's performance matches students' definitions of effective teaching), or norm-referenced (assessing the teacher's merits relative to other teachers).
- Students are generally able to distinguish between 'good' and 'bad' teachers using their own conceptions of effective teaching. They also use a variety of common criteria, evidence, and standards when responding to the different items of the evaluation form, which suggests there is a shared implicit theory of teaching among students.
- Students' ratings are influenced to some extent by their individual conceptions of good teaching.
- Students' ratings are inevitably subjective in nature, especially when the items are ambiguous or when students lack the direct relevant experience to rate the teacher on an item.

While some of the findings listed above may be plausible, Kwan (2000) recognises the limitations of his naturalistic inquiry. The number of interviewees was small and drawn from a single university in Hong Kong. In addition, the researcher's sole reliance on students' own description of their attitudes and rating behaviours may not yield accurate and reliable data, as what people say may not necessarily reflect what they think. Furthermore, data was based on a method of simulated recall, where students were

asked to think of a "good" and a "poor" lecturer and rate them before verbally reporting to the researcher on how they arrived at their ratings for each of the teachers on each rating item. This "might influence the impressions and reasoning of the students in their response" (ibid.: 42).

Nevertheless, the effect of students' conceptions of effective teaching and students' understanding of the purpose of students' ratings on their rating strategies and behaviours seem to be the theme underlying five out of the eight findings listed above. This points to the critical importance of students' role as evaluators of college teaching and highlights the argument that students' behaviours in the rating of teaching and their attitude towards the evaluation exercise is not arbitrary or uninformed. It is rather guided by their conceptions about teaching effectiveness and their perceptions of the rating process and the utility of ratings. This entails that different students may hold different understandings of what constitutes good teaching and how teachers should teach which may agree or disagree with what teachers and others in the field think. As a result, there could exist a mismatch between teachers', students', and program administrators' conceptions of effective teaching (Kwan, 2000; Kolitch & Dean, 1999; Penny, 2003; Sproule, 2000) with serious implications, especially in making personnel decisions.

A more elaborate discussion of the literature on the degree of agreement/disagreement between college students' and faculty's perceptions of the characteristics of effective teaching is presented in Chapter 3. A critical review of the literature on students' ratings, their reliability, validity, and utility is presented in Chapter 4.

2.7 Chapter Summary

This chapter presented a discussion about teaching effectiveness in higher education. The review revealed that defining the concept of effective teaching has been the subject of heated debates and contrasting views for many years. Seeds of disagreement inherent in the multidimensionality, complexity, and variability of teaching itself made reaching a conclusive definition of effective teaching an extremely difficult task. Competing theories underlying learning/teaching and the conception of teaching work have also contributed to the argument.

Despite the disagreement over the concept of teaching effectiveness evident in the literature, however, a reasonable level of agreement exists in the literature about the generic characteristics of effective college teachers. Several major reviews and many studies revealed that a number of characteristics and teaching behaviours are consistently associated with effective teachers. Some of these studies exclusively used college teachers or college students as their source of data. Other studies compared the perceptions of teachers with those of the students, reporting strong similarities between teachers' and students' perspectives about effective teaching.

The chapter also examined the research findings on the factors considered in college teachers' overall evaluations and the sources of data used in these evaluations. Evidence from North America showed that while higher education institutions today consider a wide range of criteria in making personnel decisions, such as tenure and promotion, classroom teaching effectiveness remains the most important indicator of faculty

performance. The sources of information that have traditionally been used to evaluate teaching in the last 40 years or so included systematic student ratings, self-evaluation, and classroom visitation, with student ratings becoming the most widely used source of information on teaching effectiveness towards the end of the twentieth century.

The review also identified a gap in the research literature regarding the role of the students in the evaluation of teaching. It was found that, unlike the SET instruments and their psychometric properties, the role of the students as evaluators of teaching has been largely ignored by the research in the field. Not much attention seems to have been given to how students perceive teaching effectiveness; how they view SETs; and whether their perceptions match their teachers' or not. This is one of the shortcomings in the SET research this present investigation will strive to address.

CHAPTER THREE

STUDENTS' AND TEACHERS' PERCEPTIONS OF EFFECTIVE COLLEGE TEACHING: MATCHED OR MISMATCHED PRIORITIES?

3.0 Introduction

This chapter reports on the findings of research into student and faculty perceptions of effective college teaching. The review starts with a broad view of the research literature focusing on the similarities and differences between students and college teachers in their perceptions of the characteristics of effective teachers in various contexts. After that, the review is purposely focussed to report on the results of a number of empirical studies on the subject that have been carried out in the Arab Gulf region. Given the Arab Gulf setting of this present research project, this funnelling effect is considered vital in setting out the scene and assisting the reader to better understand the contextual factors that may shape Arab Gulf students' perceptions of what constitutes good and effective teaching. It also highlights how Arab students' and faculty's differential ranking of the characteristics of effective teaching may differ from their counterparts in other parts of the world. Finally, the chapter highlights some of the implications resulting form possible mismatches between teachers and their students in their understanding of effective teaching.

3.1 Students' and Teachers' Perceptions of Effective College Teaching: Overall Correlations

As pointed out earlier in Chapter 2, many of the studies on college students' perceptions of the characteristics of effective college teaching exclusively used college students' perceptions as their data source. Other studies examined data obtained from both faculty and students. A third type of studies attempted to investigate the qualities of effective college teachers as perceived by heads of university departments. In keeping with the aims of this present study and its research questions, this section will mainly, but not exclusively, present a review of the second type of studies mentioned above. It will examine the evidence from empirical studies that have investigated the degree of match/ mismatch between college students and faculty in their perceptions of effective teaching.

Probably one of the most cited studies in the subject is that of Feldman (1988), hence the title of this chapter partially resembles the title of Feldman's review. Because of its significance as a pioneering and a benchmark study in its field, Feldman's review is cited in various sections of this chapter and its findings are discussed thoroughly. In his research synthesis, Feldman analysed thirty-one studies in each of which students and faculty at the same school or schools were asked about the importance of various instructional characteristics. The main goal of the review was to determine the extent to which students and college instructors differed in their perceptions of the importance of the various aspects of good or effective teaching. Although Feldman uses the terms "good" and "effective" interchangeably, he makes the point that in his review of the research literature, "Occasionally, "effective" teaching was more closely specified in terms of student learning" (Feldman, 1988: 292). Nevertheless, most of the studies in the

field do not seem to stress this distinction and use the term "effective" interchangeably with the term "good" (e.g. Goodwin & Stevens, 1993; Kolitch & Dean, 1999; McKeachie, 1997a), or "excellent" (e.g. Kane, Sandretto, & Heath, 2004 ;Raymond, 2001). In the present investigation the terms "good" and "effective" are used interchangeably.

After determining the differential importance of the various characteristics of effective teaching for both students and faculty in each study, the results for the two groups were correlated to indicate the degree of agreement/ disagreement between them. Feldman (1988) concluded that "Students and faculty were generally similar, though not identical, in their views, as indicated by an average correlation of +.71 between them in their valuation of various aspects of teaching" (Feldman, 1988: 291). It was also found that 12 of the 31 studies reviewed indicated correlations of at least +.85 and 9 of these 12 studies had correlations of +.90 or higher. However, Feldman cited few exceptions when the results were divided by the sample used in each study. For instance, Marques, Lane, and Dorfman (1979) reported high correlations between students and faculty in social sciences (r = +.88), the humanities (r = +.85), and engineering (r = +.80). The correlation for the natural sciences, however, was only +.16. Feldman also cited four other studies (Baum and Brown, 1980; Stevens, 1978; Stevens and Marquette, 1979; and Wotruba and Wright, 1975) involving faculty members in business schools and their students where the average correlation between students' and faculty's perceptions of effective teaching across the four studies was only +.26.

At firs glance, this may point to the possibility that the degree of match or mismatch between students and their teachers in their perceptions of effective teaching may be determined by the subject matter. However, no studies seem to have been carried out to refute or confirm this assumption. This probably reflects the complexity of such studies which may require careful control of multiple background variables, such as course difficulty, students' prior interest in the course, and assessment methods to name a few, in order to establish any causal relationship between subject matter and the level of agreement between teachers and their students on what constitutes effective teaching.

In a more recent study, Ramsden (2003) also concludes that students and lecturers are similar in their perceptions of the characterises of a good lecturer. They both identify the same set of dimensions of effective teaching: organisation, stimulation of students' interest, clear explanations, empathy with students' needs, giving feedback on students' work, setting clear goals, and encouraging independent thinking. However, examining overall correlations between students' and faculty's perceptions alone is not sufficient to identify the weights each side attaches to various specific dimensions of effective teaching. Such level of analysis requires a detailed examination of the differential ranking of these dimensions by students and their teachers.

3.2 Students' and Faculty's Differential Ranking of the Characteristics of Effective College Teaching

While overall correlations between students' and faculty's perceptions of the characteristics of effective teaching is an important indicator to be considered in identifying the degree of overall match/ mismatch between the two, examining the exact

rank order of these characteristics and their differential importance to students and faculty helps us detect the degree of importance each side attaches to various components of teaching. “The average correlation between students and teachers, while high, is not so large as to preclude some dissimilarity between them in the exact importance each group places on any particular instructional characteristic” (Feldman, 1988: 299).

Of the 31 studies included in his research synthesis, Feldman (1988) selected 18 studies for further analysis on the matches and mismatches in the differential ranking of the characteristics of effective teaching between students and faculty. Only those studies which provided sufficient information to allow the instructional characteristics to be coded into the categories used in the analysis, and to be rank-ordered in importance to students and teachers, were included. Feldman (1988) coded most of the pedagogical attitudes, behaviours, and practices of effective teaching found in these studies into 22 “instructional dimensions”. These 22 dimensions are listed in Table 3.1 overleaf along with their perceived importance for students and faculty.

Table 3.1: The Perceived Importance of Various Instructional Dimensions for Students and Faculty and Their Rank Ordering (Adapted from Feldman, 1988)

No.	Instructional Dimension	Importance* Stated by Students	Importance* Stated by Faculty	The Average Standard-ised Difference**
1.	Teacher's Stimulation of Interest in the Course and Its Subject Matter	.28 **(3.5)**	.50 **(13.5)**	+.22
2.	Teacher's enthusiasm (for Subject or for Teaching)	.32 **(5)**	.24 **(2)**	-.08
3.	Teacher's Knowledge of the Subject	.28 **(3.5)**	.23 **(1)**	-.05
4.	Teacher's Intellectual Expansiveness (and Intelligence)	.56 **(16)**	.50 **(13.5)**	-.06
5.	Teacher's Preparation; Organisation of the Course	.27 **(2)**	.28 **(4)**	+.01
6.	Clarity and Understandableness	.33 **(6)**	.39 **(5)**	+.06
7.	Teacher's Elocutionary Skills	.47 **(12)**	.57 **(19)**	+.10
8.	Teacher's Sensitivity to, and Concern with, Class Level and Progress	.22 **(1)**	.27 **(3)**	+.05
9.	Clarity of Course Objectives and Requirements	.63 **(19)**	.56 **(18)**	-.07
10.	Nature and Value of the Course Material (Including Its Usefulness and Relevance)	.46 **(11)**	.47 **(10)**	+.01
11.	Nature and Usefulness of Supplementary Materials and Teaching Aids	.54 **(15)**	.58 **(20)**	+.04
12.	Perceived Outcome or Impact of Instruction	.43 **(9)**	.51 **(15.5)**	+.08
13.	Instructor's Fairness; Impartiality of Evaluation of Students; Quality of Examinations	.45 **(10)**	.45 **(7.5)**	.00
14.	Personality Characteristics ("Personality") of the Instructor	.64 **(20)**	.70 **(21)**	+.06
15.	Nature, Quality, and Frequency of Feedback from the Teacher to Students	.49 **(13)**	.53 **(17)**	+.04
16.	Teacher's Encouragement of Questions and Discussion, and Openness to Opinions of Others	.51 **(14)**	.48 **(12)**	-.03
17.	Intellectual Challenge and Encouragement of Independent Thought (by the Teacher and the Course)	.58 **(17.5)**	.40 **(6)**	-.18

18.	Teacher's Concern and Respect for Students; Friendliness of the Teacher	.39 **(8)**	.47 **(10)**	+.08
19.	Teacher's Availability and Helpfulness	.37 **(7)**	.47 **(10)**	+.10
20.	Teacher Motivates Students to Do Their Best; High Standards of Performance Required	.58 **(17.5)**	.45 **(7.5)**	-.13
21.	Teacher's Encouragement of Self-Initiated Learning	.75 **(21)**	.51 **(15.5)**	-.24
22.	Teacher's Productivity in Research and Related Activities	.91 **(22)**	.88 **(22)**	-.03

***** This is the **average standardised rank** for each instructional dimension across the relevant studies. In order to establish comparability among the 18 studies included in the review, Feldman standardised the rank of each dimension in a study by dividing the rank of the dimension (found/ reported in that particular study) by the total number of characteristics (i.e. dimensions) included in that same study. Therefore, the smaller the fraction indicated for a dimension, the greater the rank-ordered importance of that dimension. The rank-ordered importance of each dimension is given in parentheses **(in bold)** from 1 (high) to 22 (low) for students and faculty.

****** The **average standardised difference** for each dimension was obtained by subtracting the average standardised ranks for students from the average standardised rank for faculty. A positive value indicates that students place more importance on the instructional dimension than do the teachers, whereas a negative value indicates that the teachers place more importance on the dimension than do students.

The table presented above sums up Feldman's (1988) major findings about the degree of match/ mismatch between students' and faculty's perceptions of the importance of various dimensions of college teaching. When considering the average standardised differences as a point of comparison, there appears to be three large differences in three instructional dimensions between students' and faculty's perceptions. Firstly, *stimulating students' interest* appears to receive more emphasis from students (+.22) than from faculty. Secondly, students attach less importance (-.24) than do teachers to the aspect of teachers *encouraging self-initiated learning*. Thirdly, students place less importance (-.18) than do faculty on *challenging students intellectually and encouraging independent thinking*.

Other smaller standardised differences (-.13, +.10, +.10, +.08, +.08, and -.08 respectively) also exist in the following instructional dimensions: motivating students and setting high standards of performance, teachers' elocutionary skills, availability and helpfulness to students, the perceived outcome or impact of instruction, being concerned about students, showing respect for them, and being friendly, and teachers' enthusiasm for the subject or for teaching.

Some of the differences between students' and faculty's perceptions of the importance of various instructional dimensions, however, appear more striking when they are compared by examining the rank order of the average standardised rank of every dimension for students and for faculty (Feldman, 1988). The largest differences are:

1) While intellectual challenge is ranked 17.5 for students, it is ranked 6 for faculty.

2) Stimulation of interest is ranked 3.5 for students. For faculty, it is ranked 13.5.

3) Motivating students and setting high standards ranks 17.5 for students, but only 7.5 for faculty.

Moderate differences are also reported for the following instructional dimensions:

1) Elocutionary skills are rank 12 for students and rank 19 for faculty.

2) Perceived outcome of instruction is rank 9 for students and rank 15.5 for faculty.

3) Encouragement of self-initiated learning is rank 21 for students and rank 15.5 for faculty.

4) Usefulness of supplementary materials is rank 15 for students and rank 20 for faculty.

Whether differences between students' and faculty's perceptions are determined by examining the average standardised difference or the rank-ordering of the average standardised ranks, both methods show relatively large differences for Instructional Dimensions No. 1 and No. 17. Students place greater emphasis than faculty on teachers being interesting or stimulating. On the other hand, students attach less importance than faculty on teachers being intellectually challenging (Feldman, 1988).

One of the latest studies examining the match and mismatch between students' and faculty's perceptions of effective college teachers is that of Raymond (2008). In her study, Raymond compared students' and faculty's perspectives on the importance of 11 "personality characteristics" and 14 "ability characteristics" of excellent teachers. The sample was drawn from two departments in a UAE university. Table 3.2 overleaf shows

students' and faculty's differential ranking of the importance of the 11 personality characteristics of excellent teachers:

Table 3.2: Students' and Faculty's Perspectives on the Importance of Personality Characteristics of Excellent Teachers (Adapted from Raymond, 2008)

Personality Characteristics of Excellent Faculty	Students' Ranking	Faculty's Ranking
… make classes interesting	1.5	3
… are respectful of their students	1.5	2
… are friendly to students	3	7
… care about students succeeding in their course	4	4
… show that they really like the subject they teach	5.5	5
… are fair in grading and evaluating student work	5.5	1
… are available to help students outside of class	7	8
… welcomes students' opinions/suggestions	8	6
… make an effort to get to know their students	9	10
… have a unique teaching style	10	11
… use humour in classroom	11	9

Despite the small differences in ranking, there seems to be a substantial agreement between students and faculty in their perceptions of the importance of the various personality traits of effective teachers, especially the ones pertaining to caring about students, showing respect to students, and making classes interesting.

As for the differential ranking of the "ability characteristics", the findings are summarised below in Table 3.3.

Table 3.3: Students' and Faculty's Perspectives on the Importance of Ability Characteristics of Excellent Teachers (Adapted from Raymond, 2008)

Ability Characteristics of Excellent Faculty	Students' Ranking	Faculty's Ranking
… are always well prepared and organised	1	3
… make difficult subjects easy to learn	2	4
… have many years of teaching experience	3	13

… encourage students' questions and discussions	4	1
… have expert, up-to-date knowledge of their subject	5	6
… require students to think critically	6	2
… give frequent feedback about students' progress	7	7
… expect students to become independent learners	8	5
… maintain strict control over the class	9.5	9
… encourage students to learn in pairs/groups	9.5	8
… use the latest computer technology in their teaching	11	10
… give many quizzes and tests	12	11
… lecture (talk) for the entire class period	13	14
… assign a lot of homework	14	12

Again, a high degree of agreement exists between students and faculty in their views of the importance of the various teaching ability traits listed above. Nevertheless, noticeable differences exist between students and their teachers in dimensions such as lecturer's teaching experience and requiring students to think critically. A final note to be added here is that some of the traits listed in the table above, such as *using the latest computer technology in teaching*, *giving many quizzes and tests, and lecturing (talking) for the entire class period* may not collectively fit the profile of an effective teacher in some contexts. One would assume that a TESOL teacher with access to a sophisticated multi-media lab, for instance, would not need to "talk for the entire class period", as such classes are usually characterised by independent study, self-access to information and resources, and individualised teacher support. In addition, there does not seem to be any strong research evidence to suggest that "giving many quizzes and tests" is a hallmark of excellent teachers.

As mentioned earlier, some studies in the subject used students' perspectives only as their source of data. Some of these studies are cited here, nevertheless, because they offer slightly different students' perspectives and rank-ordering from those highlighted

above. Crumbley, Henry, and Kratchman (2001) report a slightly different rank-ordering of various instructor traits that are likely to positively affect students' ratings of their instructors. These are in order of importance: fair grading (89.8%), presentation skills (89.4%), teaching style (88.8%), preparation and organisation (87.3%), and enthusiasm (82.2%). In another study by Okpala and Ellis (2005), a different set of instructional dimensions and yet a different rank-ordering for them are offered. From their analysis of data obtained from 218 U.S. college students about their perceptions of teacher quality components, the following five qualities emerged: caring for students and being concerned about their learning (89.6%), instructional skills (83.2%), subject knowledge (76.8%), dedication to teaching (75.3%), and elocutionary skills (73.9%).

Nevertheless, Feldman cautions against overlooking the similarities between students and faculty in their views about some instructional dimensions and highlights the following similarities:

1) Both students and faculty place high importance on teachers having good knowledge of the subject matter, being clear and understandable, and being sensitive to and concerned with class level and progress.

2) Both groups feel it of either moderate or moderate-to-low importance for teachers to be intellectually expansive and intelligent, and open to student questions, class discussion, and the opinions of others, and for the course material to be valuable, useful, and relevant.

3) Of distinctly low importance to students and faculty is the clarity of course objectives and requirements, the overall "personality" of the instructor, and the teacher's research activities.

However, while Feldman (1988) concludes that students and faculty were similar in their ratings of the importance of 13 of the 22 instructional dimensions he identified in his

research synthesis, he also points out that in actual rating situations students' and faculty's preferences with regard to these 13 dimensions did not always match. The following section will attempt to explore students' and faculty's perceptions of the importance of various characteristics of effective college teaching in relation to students' overall evaluation of actual teachers.

3.3 Importance of Various Instructional Dimensions Shown by Correlation with Overall Evaluations of Teachers

Surveying students' opinions on the importance they attach to various characteristics of effective teaching is not the same as asking them to rate their actual teachers. Therefore, Feldman (1988) argues that correlational analysis between students' ratings of their teachers in specific instructional dimensions and their ratings of the same teachers in global items (or overall evaluations) may be a useful tool that could: 1) offer clear evidence of the significance of specific rating items (tackling specific instructional dimensions) in deciding the overall rating of a teacher; and 2) supply indicators of the weights (or importance to effective teaching) students attach to the various dimensions of instruction. Feldman highlights the worthiness of comparing students' views of what constitutes effective teaching with their actual ratings of their teachers and says:

> At any rate, it is of some worth to compare the importance of various instructional characteristics to good teaching as determined by the views students have directly expressed on the matter (as well as by faculty views) with their "importance" as determined by the strength of their correlation with actual overall ratings of teachers. Ideally, actual ratings of teachers would be available from exactly the same students whose views about good teaching were sought.
>
> (Feldman, 1988: 315)

Unfortunately, no studies could be found that fit this ideal setting where both, students' views about the importance of various instructional characteristics and their actual

ratings of teachers, were obtained from the same sample of students. This is a gap in research that this present study is designed to fill. The alternative, but less satisfactory, procedure that most studies have followed so far is to use different students and schools for the two levels of analysis, which it can be argued, adds another source of variation and increases the complexity of the analysis.

Feldman (1976) analysed 23 studies reporting correlations between students' global ratings of their teachers and their ratings of various specific instructional dimensions of these teachers, including 18 of the 22 described in Feldman (1988) (Nos. 1-13, 15-19, as given in Table 3.1). Cross-tabulating his findings from his 1976 review with the findings from his 1988 synthesis, Feldman (1988) concluded that neither the students' nor the faculty's perceived importance of these dimensions is significantly correlated with the importance of these instructional dimensions as indicated by their power of discrimination in relation to students' actual overall ratings of their teachers. Nevertheless, it was also found that teacher's preparation and organisation, clarity, and sensitivity to, and concern with, class level and progress ranked high in importance to both, students and faculty, when asked about their views of the characteristics of effective teaching and were also of high discrimination power in the actual overall evaluations students gave to their teachers. Furthermore, it was found that teacher's ability to stimulate students' interest was highly important to students, both in theory when they give their ranking of importance and in practice when they actually rate their teachers.

Other results showed a weaker association between students' and faculty's perceptions of effective teaching and students' actual ratings of teachers. For example, both students and faculty were similar in attributing high importance to the teacher's enthusiasm and his knowledge of the subject matter. However, in actual students' evaluation of teaching, these two characteristics of instruction were only moderately important, as indicated by their correlation with the overall rating of teachers. On the other hand, teachers and students alike attached low importance to the clarity of course objectives and requirements when asked about their views of the characteristics of effective teaching, although in reality this dimension was shown to be of moderate to high significance in discriminating among teachers' global ratings by students.

As far as teacher's ability to communicate clearly with his/her students, the importance of good preparation, and teacher's enthusiasm are concerned, Feldman's findings have also been echoed by other researchers, such as Finegan & Siegfried (2000) and Ogier (2003). In an ESL setting, they found that the "communication" question of a rating form had the highest correlation with the overall rating of the teacher. Finegan & Siegfried (2000: 26), however, concluded that "the lower overall teaching effectiveness rating of ESL instructors [compared to native speakers of English] in not attributable primarily to less proficiency in spoken English but, instead, can be accounted for mostly by student perceptions of less class preparation, less enthusiasm for teaching, a less interactive teaching style, looser grading standards, and heavier reliance on multiple choice tests".

Feldman (1988), however, cautions of the limitations of correlating students' ratings of specific dimensions of teaching with the overall rating of the teacher to identify the differential importance of each dimension. He stresses that, although one would expect students' overall ratings of their teachers to be highly correlated with the instructional characteristics that they consider to be more important to effective teaching, there are a number of factors that might affect such association. According to Feldman, one factor that may affect this association is the fact that students' perceived differential weights of various instructional characteristics may simply differ from the actual weights they use when they actually rate their teachers. Another factor could be that what students consider to be highly important to good teaching "does not particularly discriminate (or, perhaps, discriminates only weakly) among teachers with respect to their overall ratings on teacher evaluation forms" (Feldman, 1988: 315).

Inconsistencies in the findings between students' perceived ranking of the importance of various characteristics of effective teaching and the weights they attach to the same dimensions of teaching in actual ratings, Feldman (1988) argues, call for more investigation. More research is needed into the effect of the setting (e.g. the type of college or academic division) or the type of students (e.g. freshmen vs. senior, or male vs. female) or teachers (e.g. the more experienced vs. the less experienced) on the degree of match/ mismatch in the criteria students and teachers use in judging good teaching. It could be also argued that other contextual factors, such as the cultural values, the prevailing pedagogical orientation of the local educational system, and the preferred learning style of individuals could also affect students' and teachers' criteria in evaluating effective teaching.

To this end, and bearing in mind the context of the present study, the following section will examine how Arab Gulf students perceive effective teaching and effective teachers as reported in a number of studies in the subject. Some of the studies included also examine the degree of similarity/difference between Arab students' and their teachers' perceptions of the characteristics of effective teaching.

3.4 Arab Gulf Students' Perceptions of Effective Teaching

Prior to discussing Arab Gulf students' perceptions of effective teaching, it is important to highlight the cultural and contextual factors that may come into play when students in this part of the world formulate their views of effective teaching and effective teachers. Raymond (2008) points out a number of "traits" and factors that may contribute to Arab students' views and evaluation of teaching. The first of these is the religion of the majority of students in the Middle East, Islam. Quoting Maamouri (1998) and Valdes (1986), Raymond (2008) brings forward the argument highlighting Islam's "rigidity" in preserving the purity and unity of its holy book, Quran, and the teachings of the Prophet Mohammed from any variation or alteration. This in turn, Maamouri and Valdes believe, dictates a "rigid" philosophy of life and a paternalistic and authoritative social hierarchy that promotes replication, imitation, rote learning, and memorisation. This philosophy, it is argued, ultimately spills out into the schools and educational systems of Arab societies.

However, in view of the prominence of Muslim scholars' significant contributions to science and knowledge throughout history, it is difficult to draw a direct causal relationship between the two without proper empirical evidence. In reality, from the

scarce empirical evidence available, results point to the opposite direction. For instance, in his study of a group of students in a UAE university, Russell (2004: 1) concluded that his sample of Arab students showed "strong beliefs and preference for deep learning approaches in addition to surface learning approaches". Also, these broad statements about Arab students may have been based on observations of the pedagogical practices in some subjects only and, therefore, cannot be overgeneralised to other subjects or teaching situations. As Raymond (2008: 57) puts it:

> These broad statements […] are perhaps more accurate of a history long past, and are perhaps more specific to how Islam and mathematics were previously taught in schools in the Gulf region, before the discovery of oil modernized society and educational systems.

The second of the "traits" or factors cited by Raymond is Arab students' preferred learning style. Citing Farquharson (1989) and Parker (1986), Raymond notes that auditory learning is Arab students' preferred style. Raymond (2008), however, contests the claim about Arab students' preference to work individually forwarded in Farquharson (1989) and cites a number of studies (Radford, 1980; Raymond, 2001; Saafin, 2005) in which Arab Gulf students were found to welcome pair or group work. Raymond's (2008) perception of Arab students, nevertheless, identifies with a category of learners described by Lowman (1995) "as anxious dependent students characterised as having excessive concern about grades and as wanting to learn exactly what the teacher wants them to learn" (p.58).

Perhaps the most important "trait" to be observed carefully in Arab students, especially by their teachers, is these students' need for, and valuation of, respect from their teachers (Raymond, 2001; Raymond, 2008; Saafin, 2005). As Raymond (2008: 58-59) puts it:

> Perhaps the most important need in the eyes of Arab students from their teachers is respect- for themselves, their culture, their country, customs and especially their religion … . The issue of "respect" is perhaps the one trait within the Arab culture which may most affect a student's attitude and behaviour in the classroom. Direct criticism of students by the authority (teacher) is interpreted as or connected with shame, and subsequently as loss of face in front of others.

With these "traits" in mind, the rest of this section will examine the findings of some studies that investigated Arab Gulf students' and teachers' perceptions of effective college teaching. As mentioned before, research in this field in this part of the world is extremely scarce. In his descriptive qualitative study of 136 Arab freshmen in the intensive English program of a UAE university, Saafin (2008) concluded that although teachers' instructional skills and ability to help students learn are perceived to be important, "certain human aspects of teachers and their attitudes toward their students are seen as crucial for judging their effectiveness" (Saafin, 2008: 1). In order of frequency, Saafin listed 13 qualities and practices of effective teachers as perceived by his UAE Arab student sample:

1. Treating students with respect
2. Being flexible and willing to compromise
3. Being helpful to students
4. Being friendly with students
5. Having a sense of humour
6. Helping students understand
7. Giving students the opportunity to speak and ask questions
8. Being dedicated to teaching
9. Being fair to students
10. Acting as a role model
11. Being knowledgeable of her/his subject
12. Being patient
13. Being cheerful and "smiling"

Once more, treating students with respect is the most frequently mentioned characteristic of good teaching. More than 61% of the students listed this aspect of instruction as an

important characteristic of effective teachers. This is in full agreement with Saafin's (2005) earlier study findings. Surprisingly, however, Arab students in this study indicated teacher's "flexibility and willingness to compromise" as their second most important criterion in judging their teachers' performance. On the other hand, characteristics of effective teaching such as being able to help students understand and having good knowledge of the subject matter, which rate highly in Western studies, appear to be of less importance to UAE students, ranking 6 and 11 respectively in Saafin's list.

In another study by Alweshahi, Harley, and Cook (2007) involving 84 final-year medical students from the College of Medicine and Health Sciences of Sultan Qaboos University, Oman, a 25-item questionnaire was administered to the student participants to determine their perceptions of the qualities of ideal bedside teachers. The questionnaire was constructed to elicit students' responses in two domains: communication and demographics. The "communication" domain emphasised instructional behaviours, such as providing constructive feedback, respecting patient confidentiality and encouraging critical thinking. The "demographics" domain, on the other hand, included characteristics such as gender, academic rank, and language skills. The researchers discovered that the dimensions in the "communication" domain were identified by the students as being far more important to ideal bedside teaching than the "demographics" domain. Using simple and clear language, using humour in teaching, giving constructive feedback, being approachable, and encouraging critical thinking were seen by the students as very important in deciding who is an ideal teacher.

Other studies focused on Arab students' preferences of learning styles rather than their perceptions of effective teaching. In Russell's (2004) study mentioned earlier, two adapted versions of the questionnaire *Approaches to Study Skills Inventory for Students* were used with a group of UAE university students to examine the assumption that students in the UAE's schools and universities prefer surface learning over deep learning. The questionnaires were coupled with other assessment tasks in the form of essay assignments. After analysing students' responses to the two questionnaires and the essay assignments, Russell (2004) concluded that the students in the sample demonstrated orientation to both surface learning approaches and deep learning approaches.

While Russell (2004) focused primarily on Arab Gulf students' preferred learning approaches, Smith (2006) made teachers' conceptions of teaching at a Gulf university the focal point of his study. Recognising the wide diversity in the cultures and ethnic backgrounds of the student population in higher education institutions around the world and the need to prepare academics to teach in such multi-cultural settings, Smith (2006) attempted to examine the lecturers' conceptions of teaching at a multi-cultural university in the UAE. As shown in some studies, like McCarger (1993), it is important for the teacher to use instructional methods that are familiar to the students and that demonstrate good understanding of their expectations. According to Smith, four categories of conceptions of teaching were found: syllabus transmission, syllabus comprehension, syllabus adaptation, and independent learning. The researcher argues that these four categories represent a shift from a teacher-guided approach to a student-centred approach. However, Smith recognises the emphasis the first three categories place on the

syllabus, which he adds is "a common feature of the more teacher-focused orientations found in conceptions of teaching" (Smith, 2006: 9). This emphasis on syllabus is also a feature of the "knowledge transmission" orientation to university teaching described by Kember & Gow (1994). According to McKeachie (1997b: 1219), students, as well as teachers, in this model prefer "teaching that enables them to listen passively- teaching that organises the subject matter for them and that prepares them well for tests". Ogier (2003) takes the argument further and cautions that this transmission model of teaching is being strengthened by the "excessive influence" of the "communication" dimension emphasised by many rating forms used in universities and colleges today.

Unlike the other studies cited above in this section, Raymond (2008) modelled her study after Feldman (1988) and examined both students' and faculty's perceptions of effective and ineffective college teaching and then compared them to identify the degree of match/mismatch between them. Using interviews and a questionnaire, the researcher surveyed the perceptions of faculty and students of effective and ineffective teaching in two different academic programs in a UAE university. The results indicated a high level of similarity between students' and teachers' views of what constitutes effective and ineffective teaching:

> Both faculty and students in this research … depicted the excellent university professor as someone who: (1) is respectful, (2) makes classes interesting, (3) is fair in evaluating, (4) cares about students' success, (5) shows a love for their subject, (6) is friendly, (7) encourages questions and discussions, (8) is always well prepared and organized, and (9) makes difficult subjects easy to learn.
>
> (Raymond, 2008: 2)

The findings of this study also seem to be consistent with the findings of past research in that both personality traits and ability traits were considered important to good teaching and effective teachers by students and faculty, but with more emphasis put on personality traits. Of all the personality traits discussed above, however, respect for students probably features as the most important quality of a good teacher in the eyes of students in the Arab Gulf.

3.5 Implications for the Mismatch between Students and Teachers' Perceptions of Effective Teaching

Mismatch between students' and teachers' perceptions of effective teaching may not only lead to distrust and conflict in college classrooms, but can also result in unfair and biased evaluation of teaching. Mismatch can even exist between teachers' goals and teaching strategies and the conceptions of teaching and learning emphasised on the typical students' rating form (Kolitch & Dean, 1999). Penny (2003) argues it further and adds:

> Given the present transition from teacher-centred to student-centred learning there is the likelihood of a mismatch between teachers' styles and students' preferences. So, when teachers choose to adopt strategies and activities that could enhance students' learning experiences, in response to the mandate given to higher education, some students might not value this. Teachers might be unfairly judged and could receive low ratings for choosing to adopt innovative strategies. To think then that students who simply desire to reproduce material to pass exams and students who desire to develop understanding and meaning in the learning process would use the same criteria to judge teaching effectiveness is simply absurd.
>
> (p.407)

Kwan (2000) suggests engaging students and teachers in a 'constructive dialogue' which explores the essence and characteristics of college teaching, which is predominantly distinguished by independent study (Sproule, 2000), as a way to create a common ground for students and teachers to develop shared meanings of teaching effectiveness

and, consequently, improve the credibility and acceptance of SET by both parties. Other researchers (e.g. Boyer, 1990; Cashin, 1996; Loder, 1990; Marincovich, 1999; McKeachie, 1997b; Scriven, 1981) also argue that there is a real need for higher education institutions to prepare their students as evaluators and observers and to train them in using rating forms and explain to them how they have been constructed and what each item represents, to enhance the validity and reliability of SET. Researchers like Emery, Kramer, & Tian (2003) and Marincovich (1999) even recommend introducing such training at an early stage during freshman seminars and the initial orientation period for maximum benefit.

The division over what constitutes effective teaching, however, exists not only between teachers and students, but also among teachers themselves.

> No group is more reluctant to admit that there are good teachers and bad teachers than college teachers themselves. This reluctance is often grounded in dubious assumptions about characteristics of teachers and students and about the nature of teaching and learning…These faculty members unwittingly dismiss the huge number of scholarly investigations that are sorting out effective and ineffective teaching behaviours.
>
> (Seldin, 1999: 1)

To this end, Fuhrmann & Grasha (1983) argue that assisting teachers clarify their assumptions is the first step in improving instruction. When evaluating teachers, "it is important to understand teachers' assumptions and definitions of good teaching" (Centra, 1993: 45). In practice, however, "most teachers are not even aware that they subscribe to a specific theory, and in fact many may apply different theories at different times or even within the same course or class period" (ibid.: 45). One can argue also that teachers' assumptions and definitions of effective teaching may not necessarily reflect a

sound understanding of the various dimensions of college teaching as demonstrated by research. This necessitates a stronger link between research and practice in this area through carefully designed and research-informed teacher induction and teacher professional development programs.

3.6 Chapter Summary

A growing body of literature is investigating students' and teachers' perceptions of effective teaching and the implications of any matches/mismatches between the two. One of the benchmark contributions to the research in this area is that of Feldman (1988). A number of studies followed with the main goal of determining the extent to which students and college teachers were similar or different in their perceptions of the importance of the various aspects of good or effective teaching. Many studies concluded that students and faculty were generally similar, though not identical, in their views.

Some of the important findings showed that, while students gave more emphasis to *stimulating students' interest* than their teachers did, they attached less importance than did their teachers to *encouraging self-initiated learning, challenging students intellectually, and encouraging independent thinking*. In Arabian Gulf-based research, a substantial, yet not total, agreement between students and faculty in their perceptions of the importance of various traits of effective teachers was also found. Several characteristics of effective teaching pertaining to caring about students, showing respect to students, and making classes interesting were emphasised by Arab students. Treating students with respect, however, was found to be the single most valued characteristic of effective teachers by Arab students in a number of studies. Some studies also showed

that, similar to what was found in the west, Arab students also placed less weight than their teachers did on the importance of requiring students to think critically.

When comparing students' perceived importance ranking of some dimensions of teaching with the importance of the same dimensions as indicated by their power of correlation with students' actual overall ratings of their teachers, the findings were inconsistent. Some dimensions, like teacher's preparation, organisation, clarity, and ability to stimulate students, were found to rank high in both. Other findings indicated a weaker association between students' and faculty's perceptions of effective teaching and students' actual ratings of teachers. For instance, while both students and faculty assigned high importance to the teacher's enthusiasm and his knowledge of the subject matter, actual students' evaluation of teaching showed only moderate correlation for these two traits with the overall rating of teachers.

With specific reference to the Arabian Gulf, some of the research in the area highlighted the cultural and contextual factors that may come into play when students in this region rate effective teaching. Religion, culture and preferred learning styles were indicated as potential factors that may influence students' conceptions of what constitutes effective teaching. Certain teacher characteristics, such as flexibility and willingness to compromise were indicated by Arab students in some studies as important to effective teaching.

Unlike the validity, reliability, and utility of students' ratings, very little research has been carried out on how students perceive effective teaching and what influences

students' evaluation of teaching. There is an urgent need for more research in this area. With the findings above in mind, it can be argued that any significant mismatches between students' and teachers' perceptions of effective teaching may potentially have serious implications on college teaching. Such mismatches may contribute to the conflict between teachers and students found in some classrooms and may even bias students' evaluations of teachers. Some researchers suggested engaging students and teachers in a 'constructive dialogue' so as to create a common ground for them to develop shared meanings of teaching effectiveness. Other researchers also call on higher education institutions to prepare and train their students as evaluators and observers of teaching if students' ratings are to be taken as an important source of information in the evaluation of teaching. It is hoped that this would ultimately improve the credibility and utility of students' ratings.

CHAPTER FOUR

STUDENTS' EVALUATION OF TEACHING (SET) IN HIGHER EDUCATION

4.0 Introduction

The previous two chapters presented a review of the literature on teaching effectiveness in higher education and students and teachers' perceptions of effective teaching. One of the underlying themes has been that effective teaching is important for good student learning and is increasingly becoming an important factor in the overall evaluation of teachers in many parts of the world. The question remains, however, how can we evaluate and improve the quality of teaching in order to improve student learning? As hinted at earlier, a growing body of literature points to students evaluations of teaching as an important source of data in evaluating the quality of teaching. "Good teaching and good learning are linked through the students' experiences of what we do. It follows that we cannot teach better unless we are able to see what we are doing from their point of view" (Ramsden, 2003: 84).

This Chapter will attempt to critically review the literature on students' evaluation of teachers (SET) in higher education from Rashdall (1936) to the present. Bearing in mind the huge body of literature on the subject and the limited scope and focused objectives of this Chapter, the review will focus mainly on the major conclusions of some of the most cited authorities in the field. The Chapter will start by examining the history and background of SET. Then, the multidimensionality of SET and the factors commonly

found in teacher evaluation forms will be discussed. After that, the reliability, generalisability, and validity of SET will be examined. Following this, there will be two sections presented to elaborate on the discussion concerning the validity and utility of SET: one will address the sources of bias in student evaluations of their teachers, while the second will present the arguments for and against the use of SET in the evaluation of faculty in higher education institutions.

4.1 History and Background of SET

Rashdall (1936), cited in Centra (1993), dates the first formal student evaluation of teachers to medieval Europe where committees of students appointed by the rector evaluated their teachers for their adherence to an agreed schedule of topics and reported any irregularities to the rector, who fined the professor for each day he had fallen behind.

According to Centra (1993), the modern history of student evaluations can be divided into four periods. The first period is from 1930 to 1960. In this period, most of the research on SET was carried out by Herman Remmers and his team at Purdue University. Following his publication of the first student evaluation form in 1927, Remmers conducted a number of studies on the subject in the period extending from 1930s to early 1960s. Some of his studies investigated the relationship of students' grades to their ratings of their teachers. Part of his work also researched the reliability and construct validity of student ratings. According to Marsh (1987), Remmers "was the first to recognize that the reliability of student ratings should be based on agreement among different students of the same teacher" and "published the first factor analysis" of

students' ratings (p.258). Because he "initiated the first systematic research program in this field", Remmers "might be noted as the father of research into students' evaluations of teaching effectiveness" (ibid.: 257).

The second period, the 1960s, marks the real start of the use of student evaluations in colleges and universities.

> The student protests that rocked so many campuses in the last half of the decade were in reaction not only to the Vietnam War and related national policies but also to policies in effect on their campuses. An irrelevant curriculum and uninspired teachers were among frequently heard student complaints. Increasingly, students saw themselves as consumers. They demanded a voice in governance ... . Evaluating their courses and teachers was one way to make their voices heard.
>
> (Centra, 1993: 49-50)

Centra himself was given the task of developing a student evaluation form at Michigan State University in 1964. However, faculty members at that early stage usually volunteered to use the forms and administrative use of the results of the ratings was infrequent due to the fact that few universities had centrally administered student rating systems (Centra, 1993). Besides the demand for accountability, students in the 60s also used SET results for course selection purposes (Onwuegbuzie, Witcher, Collins, Filer, Wiedmaier, & Moore, 2007).

The third period of SET is the 1970s. Centra calls this decade the golden age of research on student evaluation. Many studies were conducted which investigated a number of pertinent issues, such as the validity of the ratings, their biases, and their utility. Hence, many of the important studies cited in this present research were carried out in the 1970s. This decade also witnessed a wider spread of the use of SET in higher education

institutions compared to the 1960s and soon became recognised by heads of departments as the most important source of information on teaching effectiveness (Centra, 1977, cited in Centra, 1993). However, in the 1970s, SET ratings were used mainly for formative purposes and to identify the faculty development needs (Onwuegbuzie *et al.*, 2007).

During the fourth period, from the early 1980s to the present, research on SET underwent a continuous process of refinement through a series of reviews and meta-analyses. Today,

> …while commitment to teaching quality certainly varies across institutions, every college or university does evaluate teaching in some way. More and more, higher education's various publics (students, parents, legislators, and others) are insisting that we pay more than lip service to this commitment, that teaching be evaluated seriously and substantively. The time has come for higher education to put its actions where its rhetoric is.
>
> (Cashin, 1989: 1)

In the UK, student feedback questionnaires and student satisfaction surveys are widely used for the purpose of quality assurance of modules and programs (Coffey & Gibbs, 2001; Harvey, 2003; Leckey & Neill, 2001; Richardson, 2005). Nevertheless, the use of students' ratings of teacher's performance for appraisal purposes has waned recently in the UK, "with probably fewer than one-tenth of all universities doing this on a systematic basis" (Harvey, 2003: 16). Compared to the USA and Australia, "where SET data often inform decisions on tenure and promotion" (Rowley, 2003:143), using SET results for making personnel decisions, such as pay, in the UK does not seem to enjoy a lot of support. Harvey (2003) puts it strongly and says:

> There have been suggestions that student appraisal of teachers should inform performance-related pay. This is so ill-thought through as to warrant no further comment…
>
> (p.17)

Nevertheless, the increasing importance of teacher evaluation in higher education, driven by the legitimate requirements and demands of its stakeholders for effectiveness and quality of delivery and product, is quickly moving performance appraisal in colleges and universities closer to the business model of accountability and quality management. According to Darling-Hammond *et al.* (1983), the drive for accountability in education has shifted from broad issues of finance and program management to specific matters, such as the quality of classroom teaching and teacher effectiveness. Thus, in the 1980s and 1990s, SET ratings "were used mainly for administrative purposes rather than for student or faculty improvement" (Onwuegbuzie *et al.*, 2007: 115).

With the growing use of SET, however, research has increasingly cautioned against using SET information as the *only* source of teacher evaluation data and called for the diversification of evaluation tools, especially for making personnel decisions, to reflect the multi-dimensionality of teaching. But, what aspects of teaching should be evaluated to judge the quality of teaching? The following section reviews the multiple dimensions of student evaluations of teachers and sheds light on the complex behaviours of teaching in higher education.

4.2 Multidimensionality of SET

When Remmers developed what is believed to be the first student evaluation form in 1927, he based his selection of the items that were included in the form on the advice of the experts and "selected items that experts agreed were most important to teaching and

that students were capable of observing and judging" (Centra, 1993: 54). Centra also asserts that these same principles inspired the design of several current SET forms, including the most widely used commercial forms, such as the Student Instructional Report (SIR) developed at the Educational Testing Service in the early 1970s, the Instructional Development and Effectiveness Assessment form (IDEA) developed at the Centre for Faculty Evaluation and Development at Kansas State University, and Marsh's (1982a) Students' Evaluations of Educational Quality (SEEQ).

"Many SET forms have subsequently been submitted to factor analysis, resulting in item clusters- or factors- that reflect different dimensions of teaching effectiveness, as judged by students" (Centra, 1993: 54). Cashin (1995) cites a number of factor analytic studies (Abrami & d'Apollonia, 1990; Feldman, 1976; Kulik & McKeachie, 1975; and Marsh & Dunkin, 1992) that conclude that SET forms are multidimensional and measure a number of different aspects of teaching. While some authorities in the field, such as Feldman (1976, 1983, 1984, 1987, 1988, and 1989a) group student rating items into as many as 28 dimensions, others like Braskamp and Ory (1994), Centra (1993), and Mrash (1984) identify six to nine factors commonly found in student rating forms.

As evident above, the number of factors in each instrument may vary; however, shorter SET forms often combine a number of factors in a single category. "SIR's course organization and planning factor, for example, contains both the preparation and organization of the course and clarity of course objectives categories developed by Feldman" (Centra, 1993: 54-56). Following is a list of SET dimensions combining those identified by Centra (1993), Braskamp and Ory (1994), and Marsh's (1984) SEEQ:

1. Course organization and planning
2. Clarity and communication skills
3. Teacher's enthusiasm for the subject
4. Teacher-student interaction (group interaction and individual rapport)
5. Course difficulty, workload, and breadth of coverage
6. Grading, examinations and assignments
7. Student self-rated learning

In addition to the dimensions identified above, many SET rating forms, including SEEQ, also include one or two "global" rating items. When rating instruments do have a single global item, it is usually an *instructor item* which asks students to rate their teacher's overall performance in the classroom. Rating forms which have a second global item typically designate it as a *course item,* which is designed to elicit students rating of their experience in the course as a whole (Cashin, Downey, & Sixbury, 1994).

"Although there is general agreement that student ratings are multidimensional, and that various dimensions should be used when their purpose is to improve teaching, there is disagreement about how many, or which, dimensions should be used for personnel decisions" (Cashin, 1995: 2). Some studies in the field (e.g. Abrami, 1989; Abrami & d'Apollonia, 1991; Braskamp & Ory, 1994; Cashin & Downey, 1992; Cashin *et al.*, 1994; d'Apollonia & Abrami, 1997) champion a unidimensional approach and support the use of global course or instructor items or using as few dimensions as possible to provide sufficient student evaluation data for personnel decisions. Cashin *et al.* (1994) back their unidimensional approach by their findings that global items constitute most of the variance in teaching effectiveness as measured by their criterion measure, the IDEA overall evaluation measure, and, therefore, they argue that global items may be used for summative evaluation. In an analysis of students' ratings of 17,183 classes, Cashin &

Downey (1992) found that both the global instructor item and the global course item accounted for 54% and 60% respectively of their criterion variable.

However, the 'overall, this is an effective instructor' global item, which frequently appears at the end of SET forms, "is the object of considerable debate and there is no consensus about the use of global ratings even among the strongest proponents of these instruments" (Kolitch & Dean, 1999: 38). Dilts, Haber, & Bialik (1994) also oppose the idea of using global items for the purpose of making personnel decisions, arguing that asking overall assessment questions make little sense when the remaining items provide specific evidence about specific dimensions of teaching.

Others researchers argue for a multidimensional evaluation of teaching and assert the need to include all or as many dimensions of teaching as possible in summative evaluation. Dilts *et al.* (1994), for example, believe that a questionnaire of nine items is the minimum that could be used in personnel decisions. Instead of using global type items alone, some writers (e.g. Dilts *et al.*, 1994; Marsh, 1987) have suggested that various specific dimensions of teaching should be allocated differential weights for tenure and promotion decisions. Centra (1993), however, casts some doubts over this approach and questions how this weight allocation might be done equitably. Marsh (1987) suggests either allowing the instructor to determine the weight for each factor, or weight the factors according to their correlation with student learning as proven by research. Again, Centra dismisses both ideas and argues that if instructors were given the choice to determine the weights, they would probably select factors in which they usually performed best. If the factors were weighted according to research-proven

correlations with learning, Centra also cautions that this may result in some factors receiving no weight and, as a result, will be neglected by teachers. As a corrective measure, Dilts *et al.* (1994) suggest that "the easiest weighting scheme would assign greater importance to an item the higher its associated level in either the cognitive or affective domain. Such a … scheme … is less likely to be confounded with biasing influences" (p. 55). Such a scheme, however, may fail to recognise the fact that certain courses may overemphasise certain educational outcomes typical of specific level or levels of skills within the cognitive or affective domain. Some freshman courses, for example, are likely to emphasise educational outcomes at the lower end of Bloom's taxonomy, where the teacher's organisation skills are important in helping students learn facts (Cashin *et al*, 1994).

Besides the multidimensionality of SET, the reliability and validity of students' ratings is one of the most researched aspects of students' evaluations of teaching. The following two sections discuss this aspect of SET in more detail.

4.3 Reliability of SET

The reliability of students' evaluation of college teaching appears to be less contested in the literature, especially the literature on standardised rating forms, "where the reliability of class-average SETs compares favourably with that of the best objective tests" (Marsh, 2007: 8). In this literature, the reliability of SET has been investigated in relation to their stability across time, across courses, and across instructors (Young, Delli, and Johnson, 1999). In general, "Although findings are sometimes contradictory, the weight of evidence suggests that student ratings of a given instructor are reasonably stable across

items, raters, and time periods" (Murray, Rushton, and Paunonen, 1990: 250). As Centra (1993) puts it:

> The reliability of student evaluation can be best illustrated with an adaptation of Lincoln's famous remark about fooling people. We might accurately say that, with student evaluations, instructors may fool all of the students some of the time; they may even fool some of the students all of the time; but they will not fool all of the students all of the time.
>
> (p.60)

When examined from the educational measurement literature point of view, reliability in SET may refer to consistency, stability, and generalisability of items (Cashin, 1995). However, in SET research, the main source of variability is shown to be lack of agreement among students' ratings of their teacher rather than lack of agreement among different items in the rating scale. This requires another measure of reliability different from the above and this is inter-rater reliability. For ease of presentation and clarification purposes, these types of reliabilities are discussed separately in the following sections.

4.3.1 Internal Consistency

One conceptualisation of SET reliability considers the relative level of agreement among different items in a rating scale. Under this conceptualisation, the reliability of a questionnaire is traditionally estimated by the extent of agreement among a number of items designed to measure a specific underlying construct using indicators such as Cronbach's (1951) alpha coefficient (Marsh, 2007; Miller, 1995; Raykov and Shrout, 2002). According to this conceptualisation of reliability also, criticism on the reliability of students' ratings is usually concerned with the lack of internal consistency among items constituting some poorly designed rating scales, especially in instruments which have not been subjected to proper psychometric testing, such as reliability or factor

analysis. “Poorly worded or inappropriate items will not provide useful information, while scores averaged across an ill-defined assortment of items offer no basis for knowing what is being measured” (Marsh, 2007: 321).

However, it is believed that SET instruments that are carefully developed and administered can yield high internal consistency reliabilities (Aleamoni, 1999). In the original study of SEEQ (Marsh, 1982) carried out in the USA, for example, internal consistency reliability coefficients for SEEQ scales ranged from .88 to .97. In a recent study by Balam & Shannon (2010), also carried out in the USA, SEEQ scale internal consistency reliability ranged from .66 to .88. The findings in both of these studies, however, compare favourably with the SEEQ alpha coefficients of .54 to .84 found by Watkins & Regmi (1992) in Nepal.

It is often argued that Cronbach’s alpha indices are strongly influenced by the number of items in a scale (Miller, 1995; Nunnally, 1978), suggesting that the more items included in a scale, the higher the alpha values become. Kember and Leung (2008), however, argue that this technique works on the assumption that the underlying construct being measured is unidimensional. Education and the social sciences, they assert, generally deal with complex constructs that are difficult to measure with unidimensional scales:

> There can then be a tension between fully describing a construct and achieving reliable measurements. Including all pertinent facets of a construct in a scale will result in multidimensionality, which will reduce alpha values. Restricting the number of dimensions in a scale will increase alpha values, but will mean that the scale no longer address the complexity of the construct.
>
> (p.345)

Bearing in mind the small number of items in each SEEQ scale (2 in the Assignments/Readings scale, 3 in the Assessment/Grading scale, and 4 in each of the remaining scales- excluding the Course Workload/Difficulty scale), the SEEQ alpha coefficients found in the above cited studies are good. They generally point to a good level of agreement among different items designed to measure each SEEQ factor.

4.3.2 Stability

Writers like Braskamp & Ory (1994) and Centra (1993) emphasize that stability of ratings of the same teacher using the same instrument at different times is also an important gauge of the reliability of student evaluations of teachers. In general, it was found that ratings of the same teacher tend to be similar over time (Centra, 1993; Braskamp & Ory, 1994). From as early as 1954, correlations of .87 and .89 were found between students' ranking of their teachers from one year to the next (Costin *et al.*, 1971). In a study involving 1,374 students, Overall & March (1980) found a median stability coefficient of .83 across 100 courses with a time gap of one year between the two evaluations. In a longitudinal study, Marsh & Hocevar (1991a) concluded that there were no systematic changes in ratings for a group of 195 teachers over a period of 13 years.

4.3.3 Generalisability

Generalisability, which "is concerned with how confident we can be that our data accurately reflect the instructor's *general* teaching effectiveness, not just how effective he or she was in that particular course that term" (Cashin, 1995: 2) is also found to be an important indicator of SET reliability. A number of studies (Marsh, 1982b; Gillmore,

Kane, & Naccarato, 1978; and Hogan, 1973) have shown that the teacher, not the course, is the main determinant of the student rating of teaching effectiveness. A more recent study examining this aspect of the reliability of SET was conducted by Barnes and Barnes (1993). Nesting courses within instructors and instructors within courses, respectively, these researchers "found that certain instructor behaviours were resilient to variations across courses but that course evaluations were subject to specific instructors" (Young *et al.*, 1999: 181).

4.3.4 Inter-rater Reliability

Because variability in students' ratings is mainly caused by the lack of agreement between different students' evaluations of the same teacher, rather than by the inconsistency of ratings by individual students, the use of coefficient alpha does not yield adequate basis for measuring the reliability of SET instruments (Marsh, 2007). Instead, the reliability of SET responses is most appropriately measured by examining the degree of agreement among students within the same course rating the same teacher on the same dimension of teaching, i.e. inter-rater reliability (Cashin, 1995; Marsh, 2007; Marsh & Dunkin, 1992; Marsh & Roche, 1993; Penny, 2004). According to this conceptualisation of reliability, "a reliable item is one in which there is agreement among ratings within each class, but consistent differences between the ratings of different classes" (Marsh, 1982a: 81).

This alternative is a more widely used measure of SET reliability and usually considers the agreement among ratings within a given class and between classes using intraclass correlation. It is proven that the bigger the number of raters in a class, the higher the

inter-rater reliability coefficients are (Cashin, 1995; Marsh, 2007; Penny, 2004). Combining the research findings on the subject reported by Centra (1993), Marsh (2007), Marsh & Cheng (2008), and Sixbury & Cashin (1995), the inter-rater reliability coefficients for the three widely used SET instruments (mentioned under section 4.2) are presented in Table 4.1 below.

Table 4.1: Inter-rater Reliability of Three Widely Used Student Evaluation Instruments

Number of Student Raters	Inter-rater Reliability Coefficients		
	SIR One Item Overall Teacher Rating	***SEEQ*** Average for 9 Factors	***IDEA*** Median for 38 items
5	.65	.60	-
10	.78	.74	.69
15	-	-	.83
20	-	-	.83
25	.90	.90	-
30	-	-	.88
40	-	-	.91
50	.95	.95	-

Cashin (1995) reports that similar or higher reliability coefficients are usually found with well-designed forms constructed with the help of experts in the field. Cashin, however, recommends that items with fewer than ten raters or reliabilities below .70 be interpreted with caution. As Centra (1993) argues, an acceptable reliability estimate should be above .70. To overcome the problem of potential low reliabilities from small classes, Marsh & Cheng (2008) recommend averaging ratings from several small classes. For personnel decisions, such as tenure and promotion, however, Centra (1993) recommends that both the number of raters in a class and the number of courses being rated be taken into consideration. Citing research by Gilmore *et al.* (1978), Centra

(1993) also recommends sampling ratings from different course types when the purpose of collecting students' ratings is making administrative decisions, as most teachers are not likely to be equally effective in all course types.

4.4 Validity of SET

The validity of students' ratings has undergone rigorous scrutiny and debate (Marsh & Roche, 1997). "Unlike reliability, which is a necessary but insufficient condition for assessing student evaluations, validity focuses on the utility of student evaluation. Validity assesses the degree to which student evaluations of teaching performance in the classroom setting reflect actual teaching performance as exhibited by a faculty member" (Young *et al.*, 1999: 181). To this end, a very critical question about student evaluations of teachers is whether they are valid: whether they actually measure teaching effectiveness (Cohen, 1981). However, the answer to this question may not be straightforward and can be difficult to reach, as validity in SET is much more difficult to assess than reliability. A student evaluation form which consistently results in scores with adequate reliability coefficients, does not by default yield valid scores, because evidence of score reliability, although essential, is not sufficient for establishing evidence of score validity (Crocker & Algina, 1986; Onwuegbuzie & Daniel, 2002, 2004; Young *et al.*, 1999).

Nevertheless, the growing popularity of student ratings as measures of teaching quality consequently has attracted a great deal of research on their validity (Cohen, 1981). "The heavy reliance on SETs as the primary measure of teaching effectiveness stems in part from the lack of support for the validity of any other indicators of effective teaching.

This lack of viable alternatives- rather than a bias in favor of SETs- seems to explain why SETs are used so much more widely than other indicators of effective teaching" (Marsh & Roche, 1997: 1190). While certain reviews (Aleamoni, 1999; Greenwald, 1997; Theall & Franklin, 2001) argue that there are more researchers who recognised the validity of SET than those who contested it, other researchers argue that "not only has the validity of student ratings not been substantiated, but also more current empirical evidence has shown the student evaluations are misleading and/or invalid" (Crumbley *et al.*, 2001: 197-198). Perhaps the main reason behind the difficulty in establishing the validity of student evaluation of teaching is the absence of an agreed upon definition or single criterion of "effective teaching" (Cashin, 1995; Cohen, 1981; Elton, 1984; Goodwin & Stevens, 1993; Marsh, 1987, 2007). But what is exactly measured in SET validation studies?

> In validation studies, traditionally, researchers seek to provide one or more of three types of evidences: *content-related validity* (i.e. the extent to which the items on an instrument represent the content being measured), *criterion-related validity* (i.e. the extent to which scores on an instrument are related to an independent external/criterion variable believed to measure directly the underlying attribute or behaviour), and *construct-related validity* (i.e. the extent to which an instrument can be interpreted as a meaningful measure of some characteristic or quality).
>
> (Onwuegbuzie *et al.*, 2007: 116)

4.4.1 Content-related Validity

Onwuegbuzie *et al.* (2007) divide content validity into two types: item validity and sampling validity. Item validity refers to the extent to which the item characterises measurement in the targeted content area, while sampling validity refers to the degree to which the full set of items in an instrument sample the content area in a given subject. It is often the case, however, that many higher education institutions do not attempt to articulate or promote specific characteristics or behaviours of teaching as indicators of

teaching quality (Ory & Ryan, 2001). Many SET instruments are simply *ad hoc* lists of rating items, which do not reflect the multidimensionality of effective college teaching and are not supported by any empirical research or learning/teaching theory (Marsh & Roche; 1997). In the absence of content validity and empirical and theoretical analyses, the scores resulting from such forms and any decisions based on them are likely to be flawed (d'Apollonia & Abrami, 1997; Onwuegbuzie *et al.*, 2007; Ory & Ryan, 2001; Penny, 2004). Costin *et al.* (1971) also caution that such poorly designed rating instruments may be partially responsible for teachers' resistance to the use of students' ratings.

4.4.2 Criterion-related Validity

The criterion-related validity evidence has been the strongest (Onwuegbuzie *et al.*, 2007; Theall & Franklin, 2001) and most used by SET researchers (Cohen, 1981; Marsh, 2007). According to Cohen, most researchers have used this approach to establish validity by demonstrating a relationship between student ratings and other measures of teaching effectiveness.

Common indictors of teacher competence with which student ratings have been correlated are: (1) student achievement (or learning), (2) faculty self-ratings, (3) peer ratings, (4) ratings made by administrators (or expert judges), and (5) ratings made by alumni (Aleamoni, 1999; Cohen, 1981). Of the five indicators listed above, however, most researchers in the field consider student learning as the most important criterion of teaching effectiveness.

4.4.2.1 Students' Rating Correlated with Students' Learning

Historically, researchers have attempted to demonstrate that the sections that give higher ratings to their teachers are also the sections that score higher on standardised final examinations (Cohen, 1981, Feldman, 1989a; Marsh, 2007). Between 1949 and 1980 alone, more than 40 multi-section studies were conducted to test the hypothesis that students taught by more effective teachers learnt better. These studies basically compare multiple-section courses. "In the typical study, different instructors teach different sections of the same course, using the same syllabus and textbook, and most importantly using the same external final exam" (Cashin, 1995: 3). Many of these multi-section studies have been reviewed both quantitatively and qualitatively by a number of researchers. Two of the most cited reviews were conducted by Cohen (1981) and Feldman (1989a). Using the students' grades on an external exam as the measure of student learning, they investigated the correlations between the exam grade and several student rating dimensions. The average correlations were summarised by Cashin (1995: 3) as shown in Table 4.2.

Table 4.2: Average Correlations between Several Student Rating Items and Student Learning

	Average Correlations	
Student Ratings of	**Cohen (1981)**	**Feldman (1989a)**
Achievement or learning	.47	.46
Overall course	.47	-
Overall instructor	.43	-
Teacher skill dimension	.50	-
- course preparation	-	.57
- clarity of objectives	-	.35
Teacher structure dimension	.47	-
- understandableness	-	.56
- knowledge of subject	-	.34
Teacher rapport dimension	.31	-
- availability	-	.36

- respect for students	-	.23
Teacher interaction dimension - encouraging discussion	.22 -	- .36

In a more recent meta-analysis review, d'Apollonia & Abrami (1996) found that the mean correlation coefficient with SET of general instructional skills across 43 studies was .47 (Kwan, 2000). The studies indicated above clearly show moderate correlations between student evaluations of teachers and student learning, and therefore, support the validity of SET instruments.

However, it must be emphasized that these correlations are far from perfect, "in part because many of the variables that relate to students' learning will be related to student characteristics (e.g. motivation or ability), not to instructor characteristics" (Cashin, 1995: 3). Furthermore, Abrami, d'Apolonia, and Cohen (1990) identified four types of variables that might influence the correlation coefficients of multi-section students' ratings and achievements studies (p.226):

1. *Rating variables:* the quality of the rating instrument and the manner in which the evaluation takes place, such as timing and anonymity
2. *Achievement variables:* the general structure and quality of the achievement measure and the manner in which it was administered
3. *Explanatory variables:* course, student, or instruction features that may affect either the rating or the achievement measures- such as student ability, subject area of the course, instructor autonomy, or instructional setting differentially
4. *Miscellaneous variables:* methodological and other factors that might affect validity, such as the number of sections and restriction in range of scores for ratings or achievement.

4.4.2.2 Students' Rating Correlated with Instructor's Self-rating and the Ratings of Others

As mentioned earlier, student evaluations of teachers are believed to positively correlate with other indicators (other than student learning) of effective teaching, such as

instructor's self-rating, peer ratings, administrator (or trained expert) ratings, alumni ratings, etc. Feldman's (1989b) meta-analysis of 19 studies which correlated instructor's self ratings with student ratings show an average correlation of .29. In a study involving ratings of more than 50 teachers, Marsh (1987) found that the correlation between student ratings and instructor self-ratings are even higher, ranging from .45 to .62. In another study, Marsh & Dunkin (1992) asked instructors to rate two different courses in order to see whether the course rated higher by the tutor was also rated higher by the students. The median correlation based on nine factor scores between the instructor's self ratings and the students' ratings was .45.

Student evaluations of teachers have also been found to correlate with administrator's and peer's ratings. Kulik & McKeachie (1975) found an average correlation between .47 and .62 between students' and administrators' ratings. Feldman (1989b) used global items only and found a lower correlation of .39. Positive correlations have also been found between student ratings of teachers and peer ratings. Kulik & McKeachie (1975) reported average correlations of .48 to .69. Feldman (1989b) found an average of .55. Administrators' and colleagues' ratings based on classroom observation, however, have been criticised by a number of writers (Centra, 1975; Marsh & Dunkin, 1992; Scriven, 1987, 1988) because of low reliability, especially when observation is based on style and rapport only. However, when classroom observation is carried out by trained observers, Murray (1983) found out that average correlations with students' ratings tend to be higher, up to .76.

As for correlations with alumni, Overall & Marsh (1980) and Braskamp & Ory (1994) found correlations ranging from .40 to .75 between alumni ratings and current students' ratings. Feldman (1989b), on the other hand, reported a correlation of .69, belying the conventional wisdom that students will come to appreciate their teachers after they leave college and join the labour market, when they are more mature and can assess their experiences more objectively (Cashin, 1995).

Many researchers (e.g. Cohen, 1981; Marsh, 1987; Penny, 2003; Scriven, 1981), however, stress that because there is not a universal definition of effective teaching, the criterion-related approach is at best limited. Scriven's (1981) position is particularly clear in that perceived student achievement may result from a number of factors besides effective teaching, such as pressure from teachers and unreliable and/or invalid tests. Scriven also warns about the weaknesses in expert ratings based on class visits and the shortcomings of alumni surveys. He argues that expert visits to the classroom may alter teaching and yield unreliable and biased measures of teaching effectiveness while alumni ratings may be dated and out of touch with the current developments in the field. Given these difficulties, most SET researchers advocate an approach based on construct validation, where ratings are correlated with a number of teaching effectiveness criteria and uncorrelated with factors believed to be irrelevant to teaching effectiveness.

4.4.3 Construct-related Validity

An important aspect of student ratings' construct-related validity is substantive validity (Kishor, 1995; Kwan, 2000; Onwuegbuzie *et al.*, 2007; Ory & Ryan, 2001). In SET research, substantive validity assesses the degree to which the rating process used by the

students when responding to SET items is consistent with the construct being measured (Ory & Ryan, 2001). Kishor (1995) argues that if we know how students reach their ratings and what influences them, we could improve the reliability and validity of students' judgments of their teachers. Unfortunately, research in this area is very much lacking (Kwan, 2000; Onwuegbuzie *et al.*, 2007).

Another type of construct-related validity in SET research is structural validity. This type of validation research is concerned with identifying the dimensions or factors underlying the SET instrument and evaluating how well the factors resulting from students' ratings data correspond to the original factor structure of the instrument identified when the instrument was first designed. "Evidence of structural validity typically is obtained via exploratory factor analysis, whereby the dimensions of the measure are determined" (Onwuegbuzie *et al.*, 2007: 118). Onwuegbuzie *et al.* (2007), however, assert the need for structural validity evidence of SET rating forms to be compared with the dimensions of effective teaching identified in the existing literature. In this regard, SEEQ (Marsh, 1982a) is considered one of the most studied and validated SET instruments (McKeachie, 1994). Many studies have been carried out to replicate the factor structure of this instrument across contexts in the past 28 years (e.g. Clarkson, 1984; Coffey & Gibbs, 2001; Hayton, 1983; Lin, Watkins, & Meng, 1995; Marsh, Hau, Chung, & Siu, 1998; Penny, 2004; Watkins & Thomas, 1991).

A third type of construct-related validity examines the convergent and discriminant validity of SET scores. In rating instruments' research, convergent validity means factor scores from one instrument highly correlate with the factor scores from other

instruments hypothesised to measure the same construct (Marsh, 1986). Discriminant validity, on the other hand, refers to "scores generated from the instrument of interest being slightly but not significantly related to the scores from instruments that measure concepts theoretically and empirically related to but not the same as the construct of interest" (Onwuegbuzie *et al.*, 2007: 119). To illustrate this, three studies (Clarkson, 1984; Hayton, 1983; and Marsh, Touron, & Wheeler, 1985) employed the *applicability paradigm* introduced by Marsh (1981) to test the applicability and validity of two North American SET rating instruments, namely Marsh's SEEQ (Marsh, 1982a) and Frey's Endeavor (Frey, 1973), in Papua New Guinea, Australia, and Spain respectively. Despite the fact that the SEEQ and Endeavor instruments were independently developed and do not measure the same number of dimensions of effective teaching, Marsh's (1986) review of the applicability studies cited above concludes that there is a considerable factor overlap between the two instruments, with a one-to-one correspondence between five SEEQ factors and five Endeavor factors. He also concludes that correlations between the corresponding factors are substantial, while correlations between the non-matching factors are much smaller. Such findings "support the applicability and construct validity of the SEEQ and Endeavor when administered to university students in at least these countries" (ibid.: 472).

4.5 Possible Sources of Bias in SET

Various biases and factors unrelated to actual teacher performance have been hypothesised to affect the validity of students' ratings. Since the early days of research on SET instruments, numerous background variables influencing students' ratings have been investigated (for example, Al-Issa & Sulieman, 2007; Badri, Abdulla, Kamali,

Dodeen, 2006; Cashin, 1988, 1995; Centra, 1993; Crumbley *et al.*, 2001; Crumbley & Fliender, 2002; Emery *et al.*, 2003; Feldman, 1978, 1979, 1983, 1984, 1986, 1993; Kulik & McKeachie, 1975; Liaw & Goh, 2003; Marsh, 1984, 1987; Marsh & Roche, 1997; Murray, 1991; Wachtel, 1998). Many researchers (for example, Cashin, 1995; Centra, 1993; Kwan, 2000; Marsh, 2007; Martin, 1998) group these background variables into four categories:

- Course characteristics
- Student characteristics
- Instructor characteristics
- Administrative procedures & rating instrumentation

According to Broder & Dorfman (1994), cited in Young *et al.* (1999), these biasing factors can be either external or internal. External biases result from differences in teaching situations over which the teacher has little or no control, such as class characteristics and course characteristics. Internal biases, on the other hand, are attributed to student perceptions of the teacher and course characteristics and their impact on students' ratings of teachers. Students' motivational needs, lecturer's enthusiasm, and lecturer's ability to stimulate thinking are examples of these internal factors. Unlike external factors affecting SET, internal influences have received considerably less attention in the research literature tackling students' perceptions as evaluators of college teaching (Kwan, 2000; Young *et al.*, 1999). As stated in the introductory chapter of this thesis, one of the aims of this research project is to bridge the gap in this area of SET research by exploring the potential matches and mismatches between students' and teachers' perceptions of SET (Chapter 7).

It is worth pointing out at this stage that most of the studies researching bias in SET are correlational, rather than studies that show definite cause and effect (Centra, 1993). Under the four categories listed earlier, Table 4.3 below summarises the findings of a number of studies investigating the relationships between students rating of teachers and various background factors, which also can be possible sources of bias. The table combines the findings reported by Braskamp & Ory (1994) and also the findings of some of the studies that have been conducted after 1994 cited elsewhere in this chapter.

Table 4.3: The Factors Hypothesised to Influence Student Ratings of Teachers

Factor	Research findings (Effect)
Course Characteristics:	
a. Required/ elective	Ratings in elective are higher
b. level of course	Ratings in higher level courses tend to be higher
c. Class size	Smaller classes tend to receive higher ratings
d. Discipline	Higher ratings for humanities & arts, lower for social sciences, lowest for mathematics & sciences
e. Class time	No consistent effect
f. Workload	Challenging courses receive higher ratings
Student Characteristics:	
a. Gender	Inconsistent findings(although students tend to rate same sex instructors higher)
b. Expected grade	Students expecting high grades give higher ratings
c. GPA (Grade Point Average)	students with higher GPAs generally give higher ratings
d. Major/minor	Majors tend to give higher ratings
e. Prior interest in subject	Students with prior interest give higher ratings
f. Students' language of instruction in high school	Students whose language of instruction in high school is not English are more likely to be biased by the age, gender, nationality, and personality of their teacher
g. Personality	No meaningful relationships
Instructor Characteristics:	
a. Rank	No consistent relationship
b. Gender	No significant relationship
c. Teaching experience	No positive relationship
d. Personality	Warmth and enthusiasm are generally related to overall ratings

e. Instructor's nationality	Students' ratings are moderately affected by the nationality of their teacher
f. Research productivity	Positively but minimally correlated
Administrative Procedures & Rating Instrumentation:	
a. Timing of evaluation	Lower ratings are generally awarded in ratings administered during final exam
b. Student anonymity	Students give higher ratings when asked to identify themselves
c. Presence of teacher in the classroom	Students give higher ratings when their teacher is present in the classroom
d. Stated purpose of evaluation	Higher ratings are awarded if the stated purpose is promotion or tenure
e. Placement of items	Placing specific items before or after global items have no significant effect
f. Negative wording of items	No significant influence

Potential biases in students' ratings have been one of the most important causes of concern among teachers and researchers alike. "One need not talk with faculty very long to be aware of their concern about possible biases in student ratings- about variables [not related to teaching effectiveness] that correlate with student ratings" (Cashin, 1995: 4). Researchers in the field, however, disagree on the definition of bias. Centra (1993), for example, defines bias as anything that unduly influences a teacher's ratings, but is not under the control of the teacher and has nothing to do with the teacher's effectiveness. Marsh (1984) disagrees with this definition and argues that poor practices such as grading leniency would not be considered a bias under this definition. Instead Marsh & Dunkin (1997) stress the multidimensionality of SETs and argue that for a certain background factor to be judged as a bias, it must be demonstrated that this factor is not correlated with effective teaching. Marsh (2007) goes further and argues that some effects, which are sometimes interpreted as biases to SETs, should more appropriately be seen as evidence supporting the validity and multidimensionality of students' ratings.

Other researchers, although supportive of the multidimensionality of teaching campaigned by Marsh and some of his colleagues, are of the view that unless the hypothesised biasing factors can be demonstrated to affect the correlation between student ratings and student learning, they cannot be called biases (d'Apollonia & Abrami, 1997). This latter definition of bias, however, could be misleading when considering the finding that "the most impressive thing about studies relating class achievement to class ratings of instructors is the inconsistency of the results" (Kulik & McKeachie, 1975: 235). Despite the big number of studies that have been carried out to establish the concurrent validity of SET in correlation with student learning, researchers have failed, generally, to agree on the validity relationships between SET and student learning.

Cashin (1988), on the other hand, draws the attention to the extraneous factors that may affect a teacher's SET ratings but are out of his/her control, such as class size and students' motivation. Cashin argues that teachers may be faulted if they are less effective in larger classes or when teaching unmotivated students. Instead, Cashin (1988) suggests an even narrower definition of bias "restricting it to variables not a function of the instructor's teaching effectiveness" (p.3). "Thus, student motivation or class size might impact teaching effectiveness, but instructors should not be faulted if they were less effective teaching large classes of unmotivated students than their colleagues who were teaching small classes of motivated students" (Cashin, 1995: 4). Feldman (1998) in turn contests this definition and claims that such a definition has served to confuse the literature. Feldman instead drew a distinction between bias and unfairness in students' ratings of teachers, arguing that while he recognises that it is unfair to compare

instructors teaching classes of widely different sizes, he thinks that the unfairness lies in the differences in teaching circumstances rather than biases in students' ratings of teachers.

Regardless of any philosophical arguments over the definition of bias, a common concern among most researchers in the field is the argument that students' evaluations of teachers are often influenced by factors unrelated to teaching effectiveness. Many hypothesised biasing factors affecting SET have been dismissed as pure "myths" (Aleamoni, 1987, Aleamoni, 1999, Feldman, 1997) or "half-truths" (Feldman, 1997). Examining the findings of the hundreds of studies on SET in general, and the biasing factors affecting its results in particular, one can find individual studies that support almost any claim. At one end of the continuum, there are those who are very cautious about interpreting the evidence on the effects of these background factors on students' rating as bias (for example, Cashin, 1995; Marsh, 1987). For those at the other end of the continuum, this evidence of bias in students' ratings suggests that SETs can only be best regarded as "popularity contests" (Emery *et al.*, 2003), or "indices of consumer satisfaction" (Dowell & Neal, 1983), which may at times represent "the height of idiocy" (Daly, 2000). The majority in the middle, however, draw a more careful conclusion of their examination of the huge body of evidence accumulated by the hundreds of studies that examined the potential sources of bias in SET and consider SET as *one* source of data about teaching and must be used in combination with other sources of data, especially in making personnel decisions.

4.6 The Usefulness of Student Evaluations of Teachers

Despite the controversy over their validity, most writers agree that SET instruments remain the most widely used technique in measuring the effectiveness of teaching in higher education. Their results have been used differently by different groups- students, administrators, and faculty members. Although on some campuses, students have included the results of ratings in their guides for use in course selection, the use of ratings by administrators and faculty is more common (Kulik & McKeachie, 1975). With regard to faculty and administrators, evaluation of teaching effectiveness can serve many useful purposes. These include providing feedback and guidelines to teachers for improvement, directing teacher training and development efforts, assessing teacher performance for personnel decision-making purposes, assuring students and clients of effective classroom instruction, helping students select instructors and courses, enhancing the professional status and dignity of teachers, and promoting accountability of educational institutions (Murray, 1987). Probably the most disputed use of SET is their use by administrators in personnel decisions concerning faculty tenure, promotion and salaries. The amount of research in this area is overwhelming; however, the results are conflicting and inconclusive as a result of the use of different methodologies and statistical procedures which are often influenced by the researcher's literature basis defining teacher effectiveness (Ahmadi, Helms, & Raiszadeh, 2001). Therefore, academics have contrasting arguments regarding the usefulness of student evaluation of teachers.

4.6.1 Arguments for the Use of SET

As mentioned before, many supporters of the use of SET in colleges and universities argue that students evaluations of their teachers "tend to be statistically reliable, valid, and relatively free from bias or the need for control; probably more so than any other data used for evaluation" (Cashin, 1995: 6). As a rule of thumb, however, all of them also consider SET as only one source of information for evaluating teachers and call for a careful interpretation and use of its data. Apart from the arguments backed by the controversial research findings on the reliability and validity of SET, the arguments advocating the use of SET can be categorised under the following headings:

4.6.1.1 SET Feedback Results in Better Teaching and Learning

Much of the debate advocating the use of SET is driven by the belief that students ratings of their teachers will result in improved teaching and learning (Marsh, 1987; McKeachie, 1997a; Scriven, 1988). According to Scriven (1995), students are in a good position to evaluate their own knowledge and comprehension as well as motivation toward the subject. They can also observe, judge, and rate features that are believed to characterise good teachers, such as punctuality, enthusiasm, and involvement of students. Drawing on the findings of SET validation studies, many advocates refer to the positive and significant correlation between SET and student learning gains and other indicators of teaching merit (Marsh, 1987; Scriven, 1995).

4.6.1.2 SET Enhances Quality Assurance & Accountability

The growing perception of educational organisations as business organisations, or at least as business partners, and the emerging debate over the implementation of total

quality management in education is another driving force behind the calls for the use of "customer" feedback (Babbar, 1995; Cuthbert, 1996; Meirovich & Romar, 2006; Petridou & Sarri, 2004; Thakkar, Deshmukh, & Shastree, 2006). In the economic climate of today, universities and policy makers in higher education are forced to give serious thought to the issue of service quality for two main reasons. Firstly, because the expansion phase in higher education is over, therefore, there is strong competition for students. Secondly, because quality assurance systems in universities nowadays place more emphasis on the student experience as one of the evaluation criteria, students' voice in the running of higher education institutions became stronger (Cuthbert, 1996). "Thus, there has been a subtle power shift in the control of higher education from professors to their students" (Crumbley *et al.*, 2001: 197). Baba & Ace (1989) argue that the need for the use of SET in the future will grow even bigger because there is a widespread acceptance of the concept of accountability within the educational systems as a result of the increase in tax burdens to support public educational institutions.

The ongoing argument, however, remains whether students should be considered as "customers" or products and whether educational organisations are similar to or different from business organizations.

4.6.1.3 SET Instruments are Easy and Inexpensive to Administer

"From very beginning, student instructional rating questionnaires have been touted as a cheap and convenient means of evaluating the teaching of college and university faculty" (Emery *et al.*, 2003: 38). When compared to other measures of teaching effectiveness, such as classroom observation (be it by a colleague, an administrator, or a

trained observer), peer review of materials, parent reports, or professional activity portfolios (Peterson, 1995), student rating of teachers is the easiest and cheapest instrument to administer and score (Seldin, 1993b). SET questionnaires are usually machine-scored and, therefore are relatively inexpensive in terms of time and personnel.

4.6.1.4 SET Gives Impression of Objectivity

"Students reports are defensible sources of information about teacher performance" (Peterson, 1995: 86). Unlike other tools, SET instruments result in quantifiable feedback from students, which seems to be impartial and objective since SET results are reported in definite numbers. This "technical appearance" and utter simplicity have promoted the popularity of SET for many years (Emery *et al.*, 2003).

4.6.2 Arguments against the Use of SET

Kwan (2000) sums up the arguments against the use of SET, especially for making personnel decisions, into four arguments. Firstly, SET is an inappropriate measure of teaching effectiveness because students lack the maturity and expertise to judge the performance of their teachers. Secondly, SET instruments are biased and affected by situational factors which are irrelevant to teaching. Thirdly, SET is harmful to academic quality and standards. Finally, SET instruments usually contain items that are vague, ambiguous, and subjective. However, it is the view of the current researcher that the second and fourth arguments expressed above are more of validity concerns, and not a built-in defect in the SET systems which may result in a dysfunctional cycle. The arguments listed below will focus mainly on these perceived built-in defects and dysfunctional behaviours of SET.

4.6.2.1 The Dysfunctional Effects of SET on Academic Quality & Standards

Some researchers (e.g. Armstrong, 1988, and Buck, 1998) challenge the view that SET ratings lead to teaching effectiveness and claim that this argument has no real supporting evidence. Other critics of SET also believe that SET is nothing more than a "popularity contest that has little to do with learning" (Emery *et al.*, 2003: 38) and that "it is harmful to higher education" (Trout, 1997: 30). They argue that SET system causes professors and students to manipulate each other for grades and high ratings (Sacks, 1996). In order to improve their ratings and popularity, many teachers resort to inflate their grades and lighten the workload in their courses and assignments. In a survey of faculty members, (Ryan, Anderson, & Birchler, 1980) report that a third of the respondents admitted that they had substantially lowered the difficulty level and grading standards for their courses to obtain higher ratings. Crumbley & Fliedner (2002) also argue that student ratings may result in what they call "pander pollution" behaviour. " Pander pollution may be defined as purposeful intervention by an instructor inside and outside the classroom with the intention of increasing SET scores, which is counterproductive to the learning process" (Crumbley & Fliedner, 2002: 214). To sum up, the emerging argument in part of the SET literature is that overemphasis on the numerical results of SET surveys may result in a declining quality of teaching and scholarship and a lower respect for faculty (Haskell, 1997; Sacks, 1996).

4.6.2.2 Student Judgment Skills and the Dr. Fox Effect

One of the major arguments against allowing students to rate their teachers is the view that students' limited maturity and expertise does not qualify them to evaluate their teachers. According to critics, students, especially freshmen, cannot judge the

multidimensionality of teaching. Some studies have even shown that students were not fully aware of the implications of their ratings for teachers and administrators, which raises the question of how seriously students take this exercise (Ahmadi *et al.*, 2001; Dwinell & Higbee, 1993). Probably related to the ability of students to make well-informed judgments about their teachers' performance is also the issue of educational seduction, or Dr. Fox effect. Some studies have shown that students' ratings are more strongly influenced by the teacher's expressiveness and style than by content (Naftulin, Ware, Donnelly, 1973), "because charismatic and enthusiastic faculty can receive favourable student ratings regardless of how well they know their subject matter"(Emery *et al.*, 2003: 38).

4.6.2.3 Academic Freedom & Professional Values

Some researchers argue that seeking students' feedback on the teaching effectiveness of their teachers is a threat to academic freedom (Haskell, 1997). These researchers claim that SET restricts what a teacher says or does in class in his/her attempt to avoid controversial ideas or challenging learning activities and tasks. Williams & Ceci (1997) also support this conclusion and add that students' ratings force professors to think like politicians by seeking to avoid offending students with their views at the cost of substance and creativity.

4.7 Chapter Summary

Students' evaluations of teaching date back to medieval Europe, where committees of students appointed by the rector evaluated their teachers for their compliance with an agreed schedule of subjects. The modern history of SET research, however, can be

traced back to the 1930s and the pioneering research conducted by H.H. Remmers, the father of students' ratings as named by Marsh (1987). The 1960s marked the real start of the use of SET in colleges and universities, where students' ratings were used for accountability and course selection. The 1970s were the golden age of research on SET and were marked by many major studies on the validity of students' ratings, their biases, and their utility. In this decade, SET ratings were used mainly for formative purposes and to meet professional development needs. From the early 1980s to the present, research on SET has been the target of a series of reviews and meta-analyses.

Today, the growing importance of quality assurance in higher education and the mounting requirements and demands of its stakeholders for effective teaching, quality programs, and accountability is pushing students' evaluations of teaching to the top of the agenda of program managers. Therefore, SET ratings are increasing being used by colleges and universities around the world, not only for student or faculty improvement, but also for administrative and summative purposes, such as promotion and tenure.

As SETs are increasingly becoming common practice, research also increasingly cautioned against using results of students' ratings as the *only* source of teacher evaluation data. Almost all researchers in the field call for the diversification of evaluation tools, especially for making personnel decisions, to reflect the complexity and multidimensionality of teaching.

The multidimensionality of teaching has been the focus of many studies in the field in the last 40 years. The selection of items for the early SET instruments was largely based

on the advice of the experts and what they viewed as the most important to teaching. Many subsequent forms have been submitted to factor analysis, resulting in factors that reflected different dimensions of good teaching as judged by the students. This was taken as evidence of the construct validity and multidimensionality of SETs. Major studies and reviews often identify a number of dimensions as important constructs of effective teaching. These include course organisation and planning, clarity and communication skills, teacher's enthusiasm for the subject, teacher-student interaction, breadth of coverage, grading, examinations and assignments, and student self-rated learning. In addition to these dimensions, many SET rating forms include one or two "global" rating items for the course and the instructor. Despite the general agreement that student ratings are multidimensional and that all or most dimensions should be used when the aim is making decisions about the improvement of teaching, however, there is disagreement about which dimensions should be used for personnel decisions.

In the SET literature, the reliability of students' ratings is most appropriately investigated by examining the inter-rater reliability, as the main source of variance is shown to be lack of agreement among students' rating the same aspect of teaching for the same teacher, rather than lack of agreement among different items in the rating scale. The reliability of SET compares favourably with that of the best objective tests and is rarely contested in the literature.

Unlike reliability, researchers are divided about the validity of SET. Perhaps the main reason behind this division is the absence of an agreed upon definition of effective teaching. The criterion-related validity evidence has been the strongest and most

researched. Historically, many researchers have used this approach in multi-section studies to establish the validity of SET by examining the relationship between students' ratings and other measures of teaching effectiveness, most notably students' learning as measured by achievement tests. Many researchers now advocate construct-related validation studies. One type of construct-related validity being researched in SET is structural validity. This is concerned with identifying the factor structure underlying SET instruments using tools such exploratory factor analysis.

Various factors unrelated to teacher performance are hypothesised to bias students' ratings. These background variables are usually grouped into four categories: course characteristics, student characteristics, instructor characteristics, and administrative procedures & rating instrumentation. Most of the studies investigating bias in SET, nevertheless, have been correlational, rather than studies that indicate a causal relationship. SET researchers also disagree on the definition of bias and, therefore, many hypothesised biasing factors affecting SET have been dismissed as pure "myths" by some authorities in the field. The majority of researchers, however, approached the subject carefully and avoided polarised positions, urging the use of SET as *one* source of data about teaching that must be used in combination with other sources.

Despite the controversy over their hypothesised biases and utility, most researchers recognise that students' ratings have become the most widely used tool in evaluating teacher effectiveness in higher education. Their advantages as cheap, easy to administer, and easy to analyse instruments, and the sense of objectivity they convey, have won them an advanced position in the race with the more traditional methods of teacher

evaluation. Their abuse by some administrators as the only source of information about teaching in making personnel decisions concerning teacher tenure, promotion and pay, however, caused some researchers and teachers alike to perceive them as a serious threat to academic freedom and professional values.

CHAPTER FIVE

RESEARCH METHODOLOGY & DESIGN

5.0 Introduction

In this chapter, the methodology and research design of this study are explained. The chapter opens with a discussion of the research approach used in this study and its underlying philosophy. Following this is a section which provides a description of the research design, sampling and data collection methods and procedure in the two phases of the research: the qualitative exploratory study and the quantitative main study. This includes a description of the process of constructing and piloting the main study questionnaires developed by the researcher. It also provides a description of the standardised SET rating questionnaire, SEEQ, used to collect students' ratings. After that, a brief description of the data entry and screening process along with the statistical procedures used in data analyses is presented. The chapter closes with a description of the research ethics and access issues.

5.1 Research Methodology

At its simplest, the distinction between research methods and methodology can be viewed "in terms of *methods* as being some of the ingredients of research, whilst *methodology* provides the *reasons* for using a particular research recipe" (Clough & Nutbrown, 2007: 23). In other words, research methodology specifies the approach used by the researcher in the collection of data upon which inferences, interpretations, and predictions can be made (Cohen, Manion & Morrison, 2000).

Deciding on the research approach to be used in this the study was mainly guided by the following considerations:

1. The nature of the problem being investigated
2. Research aims and research questions
3. The Context and the cultural and social setting of the study
4. The educational and linguistic background of the participants
5. The timing of the study in relation to the academic calendar of the institutions sampled.

While some researchers have called for an expanded use of qualitative techniques in vocational and technology education (e.g. Coll & Chapman, 2000; Gregson, 1998) in line with the perceived strong shift from quantitative methodologies to qualitative and combined approaches in the past few decades, others (e.g. Brown, 1988; 2001) have implicitly promoted quantitative methods in second language program research. This research project used both qualitative and quantitative approaches to data collection and data analysis.

The use of a multi-method research design, in which both qualitative and quantitative research methods are used in a study, is commonly referred to as mixed methods (Creswell, 2003; Tashakkori & Teddlie, 1998, 2003). In many cases, this form of methodological triangulation (Bryman, 2004; Denzin & Lincoln, 2000; Hammersley, 1996; Johnson & Onwuegbuzie, 2004) is seen as a necessity rather than a luxury (Johnson & Onwuegbuzie, 2004; Salomon, 1991). This is particularly true in studying complex human behaviour like teaching and learning, or when the researcher is trying to

understand the problem under investigation from the participants' perspective and then collect quantifiable data on specific issues. This combination of methods yields both qualitative and quantitative data that may address the interest of a wide range of data users in the field, extending from policy makers, to other researchers, to policy implementers and front line practitioners (Gorard & Taylor, 2004; Kamindo, 2008). As Gorard & Taylor (2004) put it, "figures can be persuasive to policy makers whereas stories are more easily remembered and repeated for illustrative purposes" (p.7). In a mixed methodology approach, the weaknesses of one approach may be compensated by the strengths of the other. Using multiple methods also adds rigour, breadth, and depth to the investigation (Denzin & Lincoln, 2000).

However, depending on the topic under investigation, qualitative and quantitative approaches may be best suited and used in different phases of an investigation. For example, a qualitative approach may be used first in an exploratory study to generate hypotheses and understand the problem from participants' perspective in order to identify specific issues for a further confirmatory quantitative investigation (Bryman, 2004; Creswell, 2003, 2005; Hammersley, 1996; Johnson & Onwuegbuzie, 2004; Raymond, 2008; Tashakkori & Teddlie, 2003). "In the field of higher education, qualitative data can be a rich source of data both for generating and testing theory" (Conrad, 1982: 248).

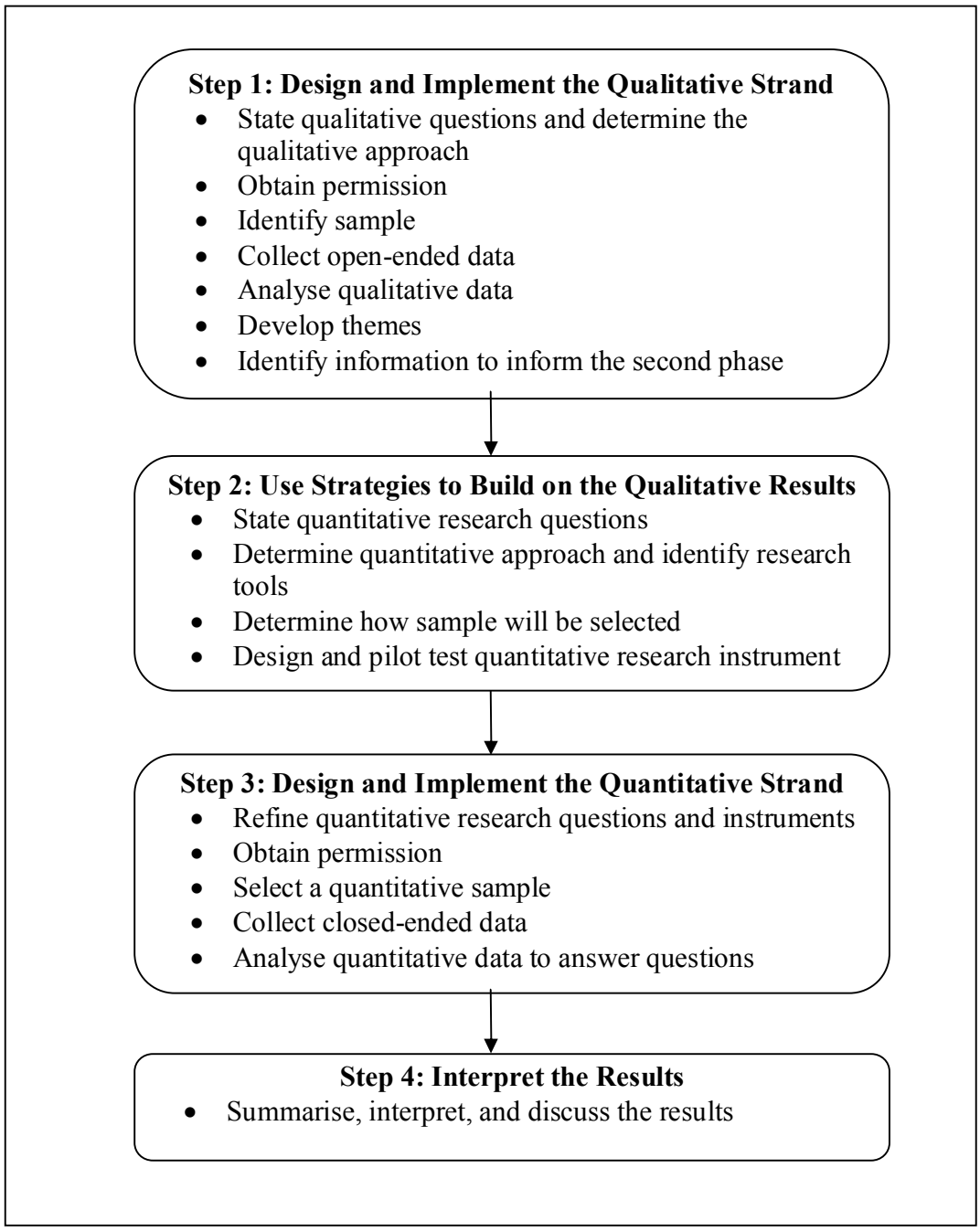
In a small scale in its qualitative exploratory phase, this study also makes use of the *comparative method* and *theoretical sampling*, two applications of the grounded theory approach (Glaser & Strauss, 1967; Glaser, 1978). "The comparative method is not built

upon a predetermined design of data collection and analysis but is a method of continually redesigning research in light of emerging concepts and interrelationships among variables" (Conrad, 1982: 243). Theoretical sampling, on the other hand, entails that "the researcher collects, analyzes, and codes his data and then decides what data to collect next and where to find them solely on the basis of the emerging theory" (ibid: 241). Conrad (1982) argues that this research strategy has strong potential for reconciling quantitative with qualitative research and theory generation with theory verification in higher education, particularly in studying areas such as college environment, impact studies, and organisational and administrative behaviour.

In the context of the present investigation, to avoid any a priori assumptions about teacher appraisal in the GFPs, and about administrators', teachers', and students' perceptions of teaching effectiveness and the characteristics of effective teachers, qualitative data were first collected from a sample of the three groups using open-ended surveys. After coding and analysing the data, major concepts were delineated and primitive hypotheses were formulated, which were then used to design further quantitative data collection and analysis. "A sequential mixed-method approach is becoming more common in research procedures as it allows the strengths of both paradigms to be made complementary, and thus provides the researcher with greater opportunity of accurately answering the research questions" (Raymond, 2008: 73).

The exploratory sequential design followed in the present study is best explained using an illustration by Creswell and Plano Clark (2011: 88) shown in Figure 5.1 overleaf with slight modifications.

Figure 5.1: The Exploratory Sequential Design (Adapted from Creswell & Plano Clark, 2011)



As Figure 5.1 shows, this investigation started exploring the problem in hand with the collection and analysis of qualitative data collected from a small sample drawn from some of the institutions that were later surveyed in the main study. In step 2, "which

represents the point of interface in mixing [methods]" (Creswell and Plano Clark, 2011: 87), the researcher used the findings of the qualitative phase to identify the salient variables, construct the research instrument for the quantitative phase, and identify relevant research tools. In the third step, the new quantitative instrument developed by the researcher along with a published standardised SET instrument, which was also quantitative in the main, were administered to a bigger and more representative sample from the GFPs. The data collected from both instruments was then analysed using descriptive and inferential statistics. Finally, in step 4, the results from the quantitative phase were interpreted and discussed.

To this end, a multi-stage, sequential mixed-model exploratory approach was used in the present investigation. Following is a description of the research methods and research design used in both stages of this research project.

5.2 Stage One: A Qualitative Exploratory Study

The aims of this e-mail based qualitative survey are:

- **To broadly explore the context and the subject under investigation in an attempt to identify pressing issues or significant gaps in perceptions between GFP administrators, teachers, and students that may require further investigation:** To avoid any a priori assumptions resulting from the researcher's familiarity with the context, it was judged necessary to survey GFP's administrators, lecturers, and students' views and perceptions about various aspects of staff appraisal, the evaluation of teaching effectiveness and characteristics of effective teachers in higher education, and students' role in the

evaluation of teaching. The study was also meant to help explore respondents' understanding of and familiarity with the topic under investigation and to weigh up respondents' attitudes and readiness to participate in a study researching staff appraisal and/or teaching evaluation.

- **To inform the focus, design and construction of the main study instruments:** This stage is designed to help narrow down the research questions and sharpen the focus of the subsequent investigation. It is also meant to help develop a conceptual framework for the study and evaluate the usefulness and relevance of the literature and documents collected for review and analysis. Depending on the emerging data and themes and the experience gained from dealing with the participants, this exploratory study is also intended to help identify the most suitable choice of questions and research tools in the main study.

5.2.1 Research Method: Open-ended Electronic Mail Surveys

It is evident from the huge body of literature on the subject that survey research is the most widely used method of data collection in the field of social sciences (Shermis & Lombard, 1999). Traditionally, researchers have relied heavily on regular mailed questionnaires to collect data from distant respondents where face-to-face interviews or other means of data collection, such as telephone or video conferencing, were not possible or feasible. "Consequently, a large body of knowledge has been generated in the various social science disciplines on the innovative ways to improve overall response rates and data quality in mailed questionnaires", but until recently very few studies have examined the use of newer 'information technologies', including email, as a way of data

collection (Mehta & Sivadas, 1995: 429). Questionnaires are defined here as "any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selecting from among existing answers" (Brown, 2001: 6).

However, parallel to the explosion in information and communication technology in the past two decades, various studies have been conducted to examine the potential and methodology of email surveys in collecting data. The major findings of this emerging body of research literature will be discussed below, and reflected upon as they apply to the current study, under four important concepts: time and cost efficiency, sample coverage, response rate, and data quality.

5.2.1.1 Time and Cost Efficiency

Email questionnaires usually have a fast turnaround time and can be sent to distant users, whether in the country where the researcher is based or abroad, quickly in almost no time (Dommeyer & Moriarty, 2000). In their review of 11 studies (Bachmann, Elfrink, Vazzana, 1996; Kiesler & Sproull, 1986; Kittleson, 1995; Mehta & Sivadas, 1995; Oppermann, 1995; Schaefer & Dillman, 1998; Schult & Totten, 1994; Sproul 1986; Tse, 1998; Tse, Tse, Yin, Ting, Yi, Yee & Hong, 1995; & Weible & Wallace, 1998) that investigated speed of response of email survey, Dommeyer & Moriarty (2000) report that, in all the eleven studies, the email survey was returned more quickly than the other methods of delivery it was compared with. In Mehta & Sivadas (1995), the difference in response speed between email and mail was considerably large. Calculated in the average number of days, the response speed for the email was 2.5 days, while for

conventional mail it was 21 days. In other studies, such as Tse *et al.* (1995), the difference in speed was much smaller, 8.09 for email and 9.79 for mail. In the present study, the response speed was lower than that reported in the above two studies. The average number of days the respondents took to return the email questionnaires was 12 days. This is mainly because of technical problems that affected internet service providers in the context of the study during data collection. These problems are discussed further in the section that follows.

Another strength of email surveys is low cost (Anderson & Gansneder, 1995; Mehta & Sivadas, 1995; Parker, 1992; Sproull, 1986). "Once respondents have access to a network, the marginal costs of collecting and communicating data electronically are much lower than costs of interviewing, telephoning, and sending questionnaires through the mail" (Mehta & Sivadas, 1995: 430). Mehta & Sivadas add that using email in data collection is even more feasible when the respondents are international. This was found to be true for the study in hand. As a research student of the University of Durham, the researcher used the internet service provided by the University to send out his email questionnaires to their recipients in Oman. However, sending only 70 completed pilot study conventional questionnaires from Oman to the UK via a courier cost the researcher more than £50, and this does not include additional administrative costs, such as paper, envelopes, copying, stuffing envelopes, transportation to and from the courier office, follow-up international calls, etc.

5.2.1.2 Sample Coverage

Unlike conventional mail questionnaires which can reach most people in the population, even in some rural areas, access to email surveys is restricted to those who have access to the internet and those who have experience in using email. Therefore, using email surveys to make generalisations about the general public is debatable, as email users may not be representative of the population (Dommeyer & Moriarty, 2000; Kent & Lee, 1999; Mehta & Sivadas, 1995). Nevertheless, the sampling frame of email and internet users is expanding rapidly in light of the current revolution in information and communication technology (Mehta & Sivadas, 1995). In addition, the potentials of email survey in reaching specific population groups, including college staff and students, have been established (Anderson & Gansneder, 1995; Mehta & Sivadas, 1995; Shih & Fan, 2008; Tse, 1998; Tse *et al.*, 1995).

For the present study, the sample targeted consists of administrators, lecturers, and students in colleges of technology in Oman. The staff members are provided with computers in their staff rooms and these computers are connected to an internal network as well as to the internet. As for students, there are a number of computer labs available for them in each college, which are also connected to the internal network and the internet. At the domestic level, many Omani households are nowadays connected to the internet and email use is becoming very common, especially among the young. In summary, "Oman has recognised educational technology, and its educational institutions have embraced it on a large scale. Many students use email facilities and surf the Web on a daily basis" (Al Musawi & Abdelraheem, 2004: 363). This makes the use of email

for data collection purposes form college populations in Oman a very feasible research technique. However, as Al Musawi & Abdelraheem (2004) point out, technical hitches, such as slow internet connection and interruptions to the service, remain a problem faced by educational institutions in Oman wanting to expand their utilisation of information technology.

5.2.1.3 Response Rate

One of the most often cited disadvantages of email surveys, compared with the conventional mail, is their low response rate (Dommeyer & Moriarty, 2000; Kent & Lee, 1999; Kittleson, 1995; Ranchhod & Zhou, 2001; Schult & Totten, 1994; Tse, 1998; Tse *et al.*, 1995). A number of studies comparing email surveys to conventional interviews and questionnaires, such as Sproull (1986) and Kiesler & Sproull (1986) demonstrated that "return rates for electronic surveys are comparable to or somewhat lower than those for face-to-face interview or mailed questionnaires" (Anderson & Gansneder, 1995: 34). Other studies, such as Mehta & Sivadas (1995), Oppermann (1995), and Parker (1992), on the other hand, demonstrated that email surveys have a higher response rate than conventional mail surveys.

Due to the limited body of literature on this relatively new survey technology, it is difficult to pinpoint the exact reasons behind this contradiction in findings. However, various factors for low response rate in email questionnaires have been discussed in this emerging body of research. Dommeyer & Moriarty (2000) sum them up in five main factors: a) email's lack of anonymity, b) the ease in which email messages could be ignored and discarded, c) some people's unfamiliarity with the email functions,

especially how to return an email, d) the difficulty of using prepaid incentives or make the respondent feel obliged to respond, and finally e) the flat text and unattractive format of an embedded email questionnaire.

The first factor, however, is debatable as it could also apply to face-to-face interviews, which are believed to yield higher response rates. In addition, many email account holders use alias email address names, which makes it extremely difficult for the researcher to establish the real identity of the respondent. In the present investigation, the vast majority of respondents used email addresses which did not reveal their real names. Especially among the female users of email in Oman, this practice is very common and in line with prevailing cultural customs pertaining to protecting women's identity and personal details from strangers. This culturally-induced level of anonymity probably reflected positively on the response rate from female students in this study, who accounted for almost half of the respondents.

As for the fifth factor, the lack of cosmetic features in email questionnaires, this can be true only for the questionnaires embedded in the body of the message itself, where layout and font options are very constrained. In attached email questionnaires, however, the situation is very different. There are many features in Microsoft Word which can help the researcher produce a very attractive and appealing questionnaire using a wide selection of fonts and styles. This questionnaire can be sent to the respondents as an attachment at the same speed, length, and cost as the embedded one, but with more appeal and formal image.

The questionnaire used in the present study was produced on Microsoft Word and sent to the respondents as an attachment. In addition to the appealing design features discussed above, well-designed questionnaires sent as attachments are also easier to fill out (Dommeyer & Moriarty, 2000). In an experiment involving business undergraduate students that compared an embedded email survey with an attached email survey, Dommeyer and Moriarty (2000) found that "the attached questionnaire was rated as better looking, easier to fill out, clearer looking, and better organised than the embedded survey" (p.46). However, the researchers also found that, while both forms of the questionnaire did not differ on response speed, number of item omissions, or response bias, the embedded questionnaire resulted in a significantly higher response rate than the attached survey. The latter finding contradicts the findings of an earlier experiment by Dillman, Sinclair, & Clark (1993) who reported that a respondent-friendly questionnaire improved the response rate of survey participants.

5.2.1.4 Data Quality

"For an email methodology to become feasible, it is necessary to demonstrate that the quality of data is equivalent to that of other survey methods" (Schaefer & Dillman, 1998: 381). The quality of data of email surveys has been generally found to be high, with qualities similar to, and sometimes even better than, paper-and-pencil questionnaires and face-to-face interviews (Dommeyer & Moriarty, 2000; Kiesler & Sproull, 1986; Mehta & Sivadas, 1995; Schaefer & Dillman, 1998; Shermis & Lombard, 1999; Sproull, 1986). In past research, the quality of survey data collected by email has been examined against a number of parameters, the most important of which are response completeness and response quality.

Email questionnaires have been shown to yield responses that are more complete with fewer item omissions compared to mail or fax questionnaires (Kiesler & Sproull, 1986; McMahon, Iwamoto, Massoudi, Yusuf, Stevenson, David, Chu, & Pickering, 2003; Schaefer & Dillman, 1998; Sproull, 1986). It was also found that responses to open-ended questions in email questionnaires were more complete and longer (Bachmann, Elfrink, & Vazzana, 1996; Mehta & Sivadas, 1995; Schaefer & Dillman, 1998). Schaefer & Dillman attribute this to the relative ease of typing an answer on a keyboard compared to writing it by hand.

Responses to email survey questions were also found to be of high quality and more clarifying and illuminating than responses to mail surveys (Mehta & Sivadas, 1995). According to a number of studies (Ayidiya & Mckee, 1990; Kiesler & Sproull, 1986; Sproull, 1986), email surveys also produced:

- less socially desirable responses to close-ended questions compared to mail questions or face-to-face interviews;
- more extreme and honest comments to subjective questions;
- more disclosing and clarifying responses to open-ended questions

"It is suggested that this type of candid response results from the fact that the computer effectively shields the respondent from the social context of traditional communication" (Thach, 1995: 30). Respondents can complete email questionnaires at their own convenience and pace and away from surveyors' control or influence. These advantages make email survey an ideal data collection method in studying subjects like the one in hand.

5.2.2 The Instrument

As discussed above, this part of the research was qualitative in nature and a formative stage in the development of the research instruments for the bigger scale quantitative study that followed. The open-ended questionnaire used in this phase of the research was designed to generate data that capture the characteristics of the setting and from which themes and ideas for further investigation could be developed. For the purpose of the study in hand, an open-ended question is defined as "an item in a questionnaire that requires participants to fill in an answer or provide a short answer in their own words" (Brown & Rodgers, 2002: 291). An "open-ended question can catch the authenticity, richness, depth of response, honesty and candour which … are the hallmarks of qualitative data" (Cohen *et al.*, 2000: 255). There are many other advantages for open-ended questions as compared with close-ended questions. Hong (1984) sums them up as follows:

> They allow the respondents to express themselves spontaneously, fully, and in their own language rather than through the predetermined choices of the researchers.
> (p. 98)

Qualitatively exploring a topic by using free response questions, for example, at the initial stage of questionnaire design is a well-established research strategy (Creswell & Plano Clark, 2011; Creswell, Fetters, & Ivankova, 2004; Greene, Caracelli, Graham, 1989; Morgan, 1998). Using qualitative tools such as open-ended questions to explore a topic prior to the design and construction of the main research quantitative questionnaire may help identify appropriate response categories for the close-ended questions. This design is sometimes referred to as the instrument development design (Creswell, Fetters, & Ivankova, 2004) or the quantitative follow-up design (Morgan, 1998). Brown (2001)

also points to the advantage of "unexpected answers" that open-ended questions sometimes bring. Such unexpected answers, Brown argues, provide valuable opportunity for the researcher to explore the dimensions of the problem and formulate narrower and more easily interpretable questions for future surveys.

In this study, no restrictions were imposed on the respondents regarding the length of their answers apart from the visual cues on the questionnaire represented by the size of the box allocated for the answer under each question. The answer box, however, was expandable to a much bigger size, as the questionnaire was emailed to the respondents as a Microsoft Word document attachment, allowing modifications to the size of the area allocated for answers when required.

Three versions of the questionnaire were produced: one for the GFP administrators (Appendix 3), one for the GFP lecturers (Appendix 4), and one for the GFP students (Appendix 5). The questionnaire in its three versions had three parts in addition to the covering letter. Part one contained background information questions concerning the participant's gender, college, first language, etc. Part two was the main section which contained the research questions. Part three was allocated for additional remarks. Part two of the administrators' and lecturers' versions had 14 questions each. Except for two questions and slight modifications in question cues to reflect respondent's role and point of view, the questions in part two of the administrators' version were similar to those in the lecturers' version. Part two in students' version had 11 questions, three of which were shared with the administrators and the lecturers.

Bearing in mind the exploratory nature of this study, its aims and objectives, and its early position in the sequence of phases forming the present investigation, it can be seen from the questions that they ranged widely and included: staff appraisal, the evaluation of teaching effectiveness and perceptions of effective teachers in higher education, and students' role in the evaluation of teaching. Students' version, however, focused entirely on students' perceptions of college teaching and teaching effectiveness in the GFP and their role in evaluating it. It was assumed that students would not be in a position to give their opinions on the staff appraisal system as a whole, partly because they would not have encountered it and also because it extends far beyond the evaluation of classroom teaching and includes other aspects, such as lecturers' contributions to committee work, relations with colleagues and superiors, and documentation of evidence. To avoid influencing participants' responses and to encourage participants to attempt answering all the questions regardless of the theme under which they fall, no labels or categories were assigned to any group of questions in all the three versions of the questionnaire. In the case of the administrators' and teachers' version of the questionnaire, for example, questions from all the three themes indicated above were mixed. The question categories were retained by the researcher for data analysis purposes only.

The content validity of the questionnaires was checked in consultation with fellow researchers and academics. The administrators' and lecturers' versions of the questionnaire were presented in English. The students' version, on the other hand, was presented in English with a translation into Arabic to ensure that students can understand the questions very well regardless of their level of English in the GFP. The translation of the questions from English to Arabic was carried out by the researcher. An independent

translation of these back into the source language, English, was conducted after that. The two versions in the source language were then compared to clarify or remove discrepancies in meaning and a final translation produced. Back translation (Ercikan, 1998; Warwick & Osherson, 1973; Axinn, Fricke, & Thornton, 1991; Chen, Liu, Ennis, 1997; Bernard, 2000; Overton & Van Dierman, 2003) is one of the most common techniques in translating research instruments used by researchers in cross-cultural research (Birbili, 2000).

5.2.3 Research Design

The data collection method used in this study, i.e. electronic mail (or email) survey, proved to be an effective and cost- and time-efficient means of data collection, which helped the researcher not only achieve the goals of the study in hand in a timely and efficient manner, but also gave him the opportunity to explore and experiment with a relatively new type of surveys. For the purpose of this study, email is defined as "a way for computer users to exchange messages (text, pictures, computer programs, audio, and video) between distant computer users" via networks connecting them (Mehta & Sivadas, 1995: 429). Expanding the definition above, one needs to add that email users nowadays in many countries around the world can also access their email accounts from wireless mobile devices, such as mobile phones and Personal Digital Assistants (PDAs).

Despite the advancements in information and communication technology infrastructure in Oman in the past 15 years and the increase in internet use at both the domestic and the institutional levels, the use of email for research and data collection purposes in

education is very rare in the Sultanate. However, Oman is probably not unique in this position. Even in the more developed world, although this application of email has been around since the 1970s (Kiesler & Sproul, 1986), it has not been discussed or researched widely despite its increasing popularity. The risk of coverage error has always been a concern for researchers using email as a data collection medium. The varying degrees in internet penetration in different population groups meant that certain populations are better suited to participate in email surveys than others. Writing in the late 1990s, Schaffer & Dillman (1998) described this limitation in email surveys:

> Thus far, the use of E-mail surveys has been restricted by the tendency of researchers to apply it only to such populations with nearly universal E-mail access. The risk of coverage error has prevented researchers from applying an E-mail methodology to other groups. (pp.378-379)

However, with the sharp increase in internet use in colleges and universities worldwide in the past decade, recent research has shown that for studies involving college populations in particular, the response rate difference between email and mail surveys is very small or even negligible, "suggesting that e-mail survey is reasonably comparable with mail survey for college populations" (Shih & Fan, 2008: 26). Shih and Fan add that college populations are more likely to respond to email questionnaires because of their familiarity with email technology, which has long been widely used in higher education institutions. These recent findings generally support the conclusions of earlier research, which also showed that electronic surveys are effective means of collecting data in academic, scientific, and business contexts in particular (Anderson & Gansneder, 1995). The following two sections discuss the method and the procedures used in collecting data via email for the current study.

The decisions made by the researcher regarding sampling and data collection methods used in this qualitative exploratory study were influenced by a number of factors, namely:

a) The exploratory nature of the inquiry and its objectives.
b) The logistics affecting questionnaire distribution and collection in the context of the study and the financial resources available to the researcher
c) The time scale governing the development of the instruments and data collection for the main study and, consequently;
d) The need to experiment with a less costly, but efficient and fast method of data collection.

5.2.3.1 Sampling

With the aims of this qualitative enquiry and its associated data analysis methods in mind, a relatively small convenience sample of 70 respondents was initially set as a target, specifically: 20 GFP administrators, 25 GFP lecturers, and 25 GFP students drawn from seven different colleges of technology. However, for reasons discussed under the *data collection* section that follows, a total of 51 completed questionnaires from four different colleges were returned. The distribution of this convenience sample by sub-group, gender and other background information is shown in Table 5.1 overleaf.

Table 5.1: Sample Distribution and Background Information for the Exploratory Study

	Male	Female	Other background information	
Administrators (**N= 8**; from 4 different colleges)	7	1	**Age group:**	
			41-50	3
			51-60	4
			Over 60	1
			First Language:	
			Arabic	3
			English	2
			Malayalam	2
			Tamil	1
			Nationality background:	
			Non-Omani Arab	3
			South Asian	3
			North American	2
			Job title:	
			Director of ELC	4
			Head of Section	4
			Years of Experience in ELT:	
			6-10	1
			16-20	1
			More than 20	6
Lecturers (**N=23**; from 3 different colleges)	12	11	**Age group:**	
			Under 25	1
			25-35	5
			36-45	6
			51-60	11
			First Language:	
			Arabic	6
			English	8
			Hindi	2
			Malayalam	2
			Telugu	2
			Urdu	1
			Shona	1
			Sinhala	1
			Nationality background:	
			Omani Arab	3
			Non-Omani Arab	3
			South Asian	8
			European	5
			African	1
			North American	3
			Years of Experience in ELT:	
			0-10	7

			11-20	6
			More than 20	10
Students (**N=20**; from one college)	11	9	**GFP Level:**	
			Intermediate	3
			Advanced	17
			Educational Region (pre-college schooling):	
			Muscat Governorate	17
			South Batinah Region	1
			Northern Batinah Region	1
			Southern Sharqiyah Region	1
			Type of school attended before college:	
			Public school	16
			Private school	4

In addition to the logistics affecting sampling and data collection mentioned earlier, the sampling method was also determined by a number of other practical considerations. Convenience sampling was perceived as sufficient for the purpose of this qualitative enquiry because making generalisations was not a goal for this phase of the research. In addition, the questionnaire was designed to be administered in an educational institution by email and not through personal contact with the subjects and, therefore, sampling became sensitive to other extraneous factors, such as the expected level of cooperation from the subjects and the technical problems that may affect the delivery of the questionnaire to the recipients. Examples of the latter are given in the following section.

5.2.3.2 Data Collection

At the outset of the study, the directors of the GFPs in the seven colleges of technology were contacted by telephone and/or email to ask for their permission and cooperation to conduct the study in their departments. Following this, the directors were sent another email inviting them with their heads of sections, lecturers, and students to participate in

the study and complete the questionnaires, which were attached to the message as MS Word documents. In all the three versions of the questionnaire, the covering letter included instructions directing respondents to return the completed questionnaire to the researcher using the given email address. The letter also stated the aims of the study and assured the respondents of the confidentiality of the data and the answers they provide.

The directors were requested to advertise the study in their departments by posting the covering letter of the lecturers' version of the questionnaire in their departments and by directly distributing the questionnaire to their lecturers using their internal computer networks. Two directors even prepared mailing lists for their teacher volunteers and forwarded them to the researcher, which were very useful for follow-up purposes. As for the students, the directors were also requested to inform their GFP students about the study by posting the translated covering letter attached to the students' version of the questionnaire in their departments and classrooms and by word of mouth through their teachers. Student volunteers were given the chance either to enlist their email addresses with their teachers or GFP director's office to be forwarded to the researcher, or contact the researcher directly using the email address indicated in the letter. This was the only way students could be involved, as the GFPs did not have administrator rights over students' user accounts or mailing lists. Furthermore, many students preferred to use their personal email addresses.

As shown in Table 5.2, only eight GFP administrators, from four different colleges, completed the questionnaire. As for the lecturers, 23 respondents from three colleges participated in the study. Students' sample, on the other hand, was less spread out as

only one college, out of the four that participated in this study, managed to a compile a list of student volunteers. From this college 20 students returned the questionnaire. It is not clear why none of the students from the other three colleges volunteered to take the survey. The main reason could be the Internet outage which affected Oman and various other parts of the Middle East and South Asia because of damage to undersea Internet cables in the Arabian Gulf and the Mediterranean in late February and early March 2008. The outage hit headlines in the media around the globe around the same time when the questionnaires were sent out. Some administrators and lecturers made a note of these technical problems in their correspondence with the researcher. While the differences in the technical level and capability of the internal communication infrastructure between the seven colleges are minor, it is not clear whether the reported interruptions in internet service affected some areas in the country more than others to explain the variation in response rates between the different colleges.

5.2.4 Data Analysis

After receiving the completed questionnaires from the respondents via email, answer boxes for each question were copied onto an Excel worksheet. Responses to questions ranged from extended comments to short, one or two word answers. Because of the conciseness of some responses, and because the questions within each major theme were interrelated, the focus during data analysis was shifted from individual questions to major themes of questions. Assuming there was continuity within each participant's response to the questions within each theme, the analysis moved from the part (single items) to the whole (set of items within a theme). This method of analysis for responses to open-ended surveys on perceptions of effective teaching is consistent with that of

Prat, Kelly, & Wong (1999). Where students' answers were given in Arabic, the responses were first translated into English following the same approach described in the previous section and then included with the rest of the responses for analysis. Literal translation of responses is seen to lead to a better understanding of the participants' mentality (Honig, 1997), an aspect that elegant free translation may lack.

Each response was read and reread to get a feeling of the data and identify general patterns and themes in the responses for each group of questions. These common themes were colour coded and noted next to each group of answers. Following this, response categories were developed for each general theme in order to group and label the comments in each response. On the Excel sheet, responses were entered in one column and response categories were entered in the next column. Once the categories had been identified and the responses coded for each theme, data across all the questions was examined again for relations between categories or major trends. This method is typical of the comparative approach (Glaser & Strauss, 1967; Glaser, 1978) of data analysis. This approach also draws from the phenomenological method which requires the researcher to suspend prejudgment and not impose meaning too soon (Holliday, 2002). Finally, a summary of the findings was prepared incorporating narratives from the respondents where necessary (Appendix 6).

As pointed out at the beginning of this chapter, one of the main goals of this exploratory phase of the study is to identify any gaps in perceptions between administrators, teachers, and students in the GFPs with regard to various aspects of staff appraisal, the evaluation of teaching effectiveness and characteristics of effective teachers in higher

education, and students' role in the evaluation of teaching. As shown in Appendix 6, the data generated from this exploratory study identified two major gaps in perceptions between GFP teachers and their students. The first gap concerns the perceived importance teachers and students attach to various dimensions of effective teaching. The second gap is related to how teachers and students perceive the reliability, validity, and utility of students' ratings of teachers and the role of students as evaluators of college teaching. Based on these findings, the second stage of the research was designed.

5.3 Stage Two: The Main Study

Owing to the study objectives and research questions and the findings of the exploratory study, a second phase of the investigation was deemed necessary. This second phase is designed to generate quantifiable data from a much bigger and more representative sample, which can be used to make reasonable generalisations about the subject in hand and help inform the policy and practice of teacher evaluation in higher education institutions in Oman and similar contexts worldwide.

5.3.1 Research Method: Quantitative Surveys

As stressed earlier, this phase expands on the exploratory study findings, specifically those findings that pointed to a significant gap between the GFP teachers and students in their perceptions of the characteristics of effective teachers and the role of the students in the evaluation of teaching. As explained further under the Instrumentation section, two quantitative questionnaires were used in this phase of the study to operationalise and quantify the concepts that emerged from the first stage. The first questionnaire, entitled *Perceptions of Good College Teaching & Students' Evaluation of Teachers*, was constructed by the researcher himself, while the other one was SEEQ, the widely used standardised SET instrument developed by Marsh (1982a), discussed in Chapter 4.

Likert-type scales are used in this phase as the primary tool of data collection. In the construction of the first questionnaire, data collected from the open-ended survey questions, along with the findings from the literature review, were used to create statements for a ranked scale to identify the participants' differential ranking of the importance of various characteristics of effective teachers. In the same way, another

scale with statements measuring respondents' perceptions of students' ratings of college teaching, their validity, utility, and the role of the students in teacher evaluation was produced. The SEEQ also included nine ranked scales to rate nine different dimensions of teaching, in addition to two open-ended questions and a section for additional remarks at the end of the questionnaire.

Brown (2001) lists four types of scales: nominal, ordinal, interval, and ratio scales. Nominal scales are most appropriate for quantifying constructs or variables in categories or groups, like gender or first language in the present investigation. Ordinal scales, which are sometimes called ranked scales, involve creating ordinal numbers along a scale to represent the variable being measured. Such scales include ranking scales which ask respondents to rank objects or concepts. For example, one of the ranking scales used in this phase of the investigation asks respondents to rank the importance of various characteristics of effective teachers along a scale which ranges from 1=Not at all important to 5= Extremely important. The third type of scales, interval scales, also describe the rankings along a scale, but with equal intervals between the points on the scale. "However, some questionnaire results, such as attitude scales, are also treated as interval scales" (ibid.: 18). The fourth type is ratio scales. Ratio scales, not only require a measurement along a scale with equal intervals, but also require that the ratios of values along the scale are meaningful and this entails that the scale must have a true and meaningful zero point (Field, 2009). A scale asking teachers for the number of years of teaching experience, for instance, is an example of a ratio scale.

The participants' avoidance strategies that may exist in responding to some open-ended questions can be neutralised by using scales for data collection. There are other practical considerations behind using Likert scales also. One of the advantages of scales is their suitability for generating responses that can be easily submitted to statistical analysis for comparison and ranking purposes, since variations in responses are limited. "Of all the research methods, survey research may be the most practical and usable in one sense: It relies more on common sense and less on complex statistics" (Brown, 2001: 15). Rating scales are also considered ideal for determining respondents' opinions, beliefs, attitudes and perceptions, "for they combine the opportunity for a flexible response with the ability to determine frequencies, correlations and other forms of quantitative analysis. They afford the researcher the freedom to fuse measurement with opinion, quantity and quality" (Cohen *et al.*, 2000: 253).

5.3.2 Instrumentation

The *Perceptions of Good College Teaching & Students' Evaluation of Teachers* is designed to probe the matches and mismatches between the GFP teachers and their students in their perceptions of the importance of various characteristics of effective teachers and their views about students' evaluation of college teaching. The SEEQ was administered to the same groups that completed the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire. Students' ratings obtained with SEEQ were subjected to statistical analysis to: a) determine students' ability to identify the various components of effective teaching underlying the instrument; b) measure the inter-rater reliability of students' ratings of their teachers; and c) to identify

the effect of various student, teacher, and course background variables on student ratings.

5.3.2.1 The Pilot Run

The pilot run for the instrument developed by the researcher, *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire, involved three steps: construction, testing, and refining.

Construction: The pilot *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire had two versions, one for lecturers and administrators and one for students. Again, the questionnaire was designed to be used in the GFPs in Colleges of Technology in Oman. The questionnaire consisted of two Likert Scales. The first scale asked lecturers' and students' to rank the importance of a number of characteristics of effective college teaching derived from the existing literature and the findings of the exploratory study on a four point scale: 1= Not at all important, 2= Slightly important, 3= Moderately important, and 4= Very important. The second scale investigated lecturers' and students' views and perceptions of SET and staff appraisal in general in their departments- again with themes taken from the existing research literature and the findings of the exploratory study. This scale consisted of 4 points: 1= Strongly disagree, 2= Disagree; 3= Agree, and 4= Strongly agree.

Despite the many similarities between the two versions in layout, instructions, and underlying variables being investigated, there were few differences between the two. Although the items in the first scale in the two versions were the same for both lecturers

and students, the second scale in students' version had fewer items. Some items about staff appraisal, which were considered difficult for students to answer because they required specialist knowledge of the procedures followed in staff evaluation, were confined to the teacher's version only. In addition, unlike the lecturers' version which was presented in English only, the students' version was presented in English with translation into Arabic. Similar to the open-ended survey in stage one, back-translation technique was used to translate this questionnaire from English to Arabic.

Testing: After expert review was received from the main supervisor and colleagues at Durham University and in Oman on the design and content validity of the questionnaires, a few changes were made to the scales and a covering letter explaining the purpose of the study and assuring the participants of the confidentiality of the study was produced for both versions. Then, one of the CTs was contacted and asked for permission to administer the pilot questionnaires to a small sample of its lecturers and students in the GFP. A relatively small sample of 31 lecturers and 40 students was selected for the pilot run. The decision taken at that stage was to reserve as many students and lecturers for the main study as possible, as participants in the pilot run could not be included again in the main study (Creswell & Plano Clark, 2011).

The request was approved and the questionnaires were sent to Oman with detailed sampling and administration guidelines. Stratified random sampling was used to ensure that: 1) all the levels in the GFP are represented; 2) there is a reasonable ratio of male to female; and 3) and for the teachers' version, both native and non-native speakers of English are involved. The questionnaires were administered by the management of the

Language Centre in that college using the guidelines provided by the researcher. In early March 2008 the completed questionnaires were returned to the researcher in the UK.

Data from the questionnaires were entered onto SPSS datasets. Initial calculations of the reliability of the scales using Cronbach's coefficient Alpha were performed using the software package Statistical Package for Social Sciences (SPSS) for Windows, version 15.0. A cut off point was set at .70. This is an acceptable Alpha value for measures of attitudes (Crowl, 1996), which are generally difficult to measure. The results of this analysis showed a moderate Alpha coefficient of .79 for scale one, section one, in the students' version. However, the same scale in the lecturers' version scored a relatively low coefficient of .58. Scale two in section 2, on the other hand, failed to reach the cut off point of .70, for both students and lecturers. For students, it was .45, while for lecturers it was .62.

Refining: Based on the results of the reliability analysis and the feedback given to the researcher from fellow teachers, researchers and experts in the field, a decision was made to unify the number, format, and wording of items in both versions of the questionnaire. Bearing in mind the comparative nature of several of the research questions in this study, this unification was seen as a necessary step to facilitate the analysis and interpretation of data.

Work on refining the questionnaires was not confined to improving their reliability and validity but also tackled other aspects, such as layout, wording, design, and breadth of coverage. The completed questionnaires were checked for any emerging patterns in

missing values that may be attributed to faulty questionnaire design. Also, following the analysis of the distribution of responses in both scales, response categories in scale one were increased from four to five categories to create more equal intervals.

The instructions at the beginning of each scale were also revised to incorporate some of the comments made by some respondents. Using SPSS reliability analyses, some items were deleted to improve the Alpha value for the scale, while new items were added, mainly because of splitting up some existing items. Among the items deleted were the ones probing respondents' attitudes and opinions about certain aspects of staff appraisal, such as evaluation by heads of departments and levels of satisfaction with the existing staff appraisal system in general. To aid data analysis and improve comparability of findings from the two populations, the items in scale 2, too, were unified for both students and teachers. A detailed description of the revised instrument is presented below.

5.3.2.2 The Revised version of *Perceptions of Good College Teaching and Students' Evaluation of Teachers* questionnaire

The revised versions of this questionnaire, for both students (Appendix 7) and teachers (Appendix 8), were exactly the same in terms of the number of scales and the number and wording of items in each scale. In both questionnaires, section one asked respondents to rate the (perceived) importance of 38 characteristics of effective college teaching on a scale of five points: 1= Not at all important, 2= Slightly important, 3= Moderately important, 4= Very important, and 5= Extremely important.

Section two, on the other hand, asked participants to respond to 19 items about their perceptions of students' rating of college teaching in three sub-themes: the factors hypothesised to bias SET, the utility of SET, and the role of the student in SET. This scale measured respondents' perceptions on 4 points: 1= Strongly disagree, 2= Disagree; 3= Agree, and 4= Strongly agree.

Like the pilot questionnaire, teachers' version was presented in English, as all of the teachers involved are TESOL teachers who are fluent in English. Students' revised version, on the other hand, was presented in Arabic only and without translation into English. This decision was made based on two observations from the pilot study. Firstly, the questionnaire is administered to students who are native speakers of Arabic, but with diverse levels of English proficiency, ranging from elementary to advanced. Judging by the marks and notes left by the students in the Arabic text in the pilot questionnaire, such as clarifications, explanations of responses, and underlined or circled key words, and the absence of the same from the English text, it was assumed that most students read the Arabic version of the items. Secondly, as more items were added to the revised version, inserting English translation next to the items in Arabic made the questionnaire look crammed, longer and less user friendly. It was feared that this could discourage the students from completing the questionnaire.

Another change is that the background information section in the students' version was moved to the beginning of the questionnaire, unlike the lecturer's version in which the background information section remained at the end. This is because the background information section in the students' version required students to provide data deemed

essential for data analysis required to answer several of the research questions, such as their gender, group number and level of English, which may be compromised by fatigue or anxiety to leave the class if left until the end of the questionnaire. The student's ID number was also required in this section to allow pairing this questionnaire with the SEEQ ratings for each student for potential correlational analysis. In addition to the confidentiality assurances provided in the covering letter, the researcher personally assured the students of the confidentiality of their responses and the data they provide during the administration of the questionnaire.

5.3.2.3 SEEQ

Because some of the main research questions probe the reliability and factor structure of the Western-developed but internationally used SEEQ in Oman, no piloting was carried out on this instrument prior to using it in the field. In a way, the administration of SEEQ in Oman may be considered as a pilot run and an applicability test at the same time, but with the added benefit of using a relatively big, representative sample.

SEEQ is a multidimensional standardised rating form and it is considered one of the most reliable, valid standardised rating forms for measuring teaching effectiveness (Coffey & Gibbs, 2001). The research leading to the development of the first SEEQ was originally conducted at the University of California, USA, in the 1970s and early 1980s (Marsh, 1982a). The current SEEQ was developed by Marsh (1982a) at the University of Southern California. According to Marsh and Dunkin (1992), SEEQ's items originated from various sources. Among these are extensive reviews of the literature and similar

rating forms, and interviews with university teachers and students about their perceptions of effective college teaching.

As noted in Chapter 4, SEEQ's reliability reported in the SET research literature is very high. The validity of the instrument has also been researched extensively. "The validity of the SEEQ is based on over 30 factor analyses in different settings, multitrait-multimethod analyses, logical analysis of the qualities of effective teaching, as well as being supported by principles of adult learning" (Penny, 2004: 162).

The questionnaire consists of four parts. Part 1 asks the students for demographic and background information about themselves, the course, and the lecturer. This part also includes instructions to the students on how to complete the questionnaire. Part 2 is the main part that includes the rating items. As shown in Appendix 9, the 31 rating items are grouped into nine five-point Likert scales, from 1= strongly disagree to 5= strongly agree, and measure different dimensions of teaching effectiveness as follows:

Scale 1: Learning/Academic Value (4 items): This scale consists of four items which measure whether students found the class intellectually challenging and whether their interest in the subject increased as a result of taking the course.

Scale 2: Instructor Enthusiasm (4 items): This scale asks students to rate their teacher's enthusiasm and whether he or she is capable of giving presentations that hold students' interests.

Scale 3: Organisation/ Clarity (4 items): It asks whether the course materials were well prepared and clearly presented, and course objectives adequately met.

Scale 4: Group Interaction (4 items): This scale asks students to rate their teacher's ability to encourage students to participate in class discussions and express their own ideas, and to seek help from the teacher.

Scale 5: Individual Rapport (4 items): It rates the teacher's ability to provide opportunities that take account of the individual differences between students and his/her accessibility to students seeking help and support.

Scale 6: Breadth of Coverage (4 items): It measures whether the teacher discusses various points of view and whether he or she contrasts the implications of various theories.

Scale 7: Assessment/ Grading (3 items): This scale evaluates the quality, fairness, value, and relevance to course objectives of the teacher's feedback and graded materials used by him/her.

Scale 8: Assignments/Readings (2 items): This scale asks students to rate the value of the texts and supplementary readings assigned by the teacher, and evaluate the contribution of assignments to the appreciation and understanding of the subject matter.

Scale 9: Overall Rating (2 items): These are the global rating items. One asks for the overall rating of the teacher, while the other one is for the overall rating of the course.

It is worth pointing out at this stage that the SEEQ also includes an additional factor called the Workload/ Difficulty factor. However, this factor was treated as a background factor and was not included in the factor analysis or the reliability analysis carried out on the ratings, because it was not considered a target dimension of a lecturer's teaching performance. It was considered a feature of the course itself rather than the teacher who teaches it. This decision is consistent with that of Marsh & Roche (1993) in their investigation of the applicability of the SEEQ in an Australian setting. It is also consistent with the instrument developer's view that SEEQ ratings are primarily an evaluation of the teaching performance and characteristics of the person who teaches the course, rather than of the course itself (Marsh, 1981, 1982b).

The decision was also informed by the findings of previous SEEQ research on the discrimination power of the items on this factor. In a number of studies investigating the applicability of the SEEQ in different countries, it was found that the Workload/Difficulty items did not differentiate between good and poor instructors (Hayton, 1983; Clarkson, 1984; Marsh, Touron, & Wheeler, 1985; Marsh, 1986). Furthermore, the Workload/Difficulty factor contained background variables (e.g. perceived course difficulty, workload, and expected grade) that have long been investigated as possible sources of bias in students' ratings. Because one of the research questions in the present study (Question 7) was designed to test for associations between these background variables and students' overall ratings, it was considered important to

separate the Workload/Difficulty background items from the rest of the rating items. This position is in agreement with the developer of SEEQ who advises that such variables "not be included both as items on which students rate teaching effectiveness and as background characteristics, particularly when reporting some summary measure of variance explained" (Marsh, 1984: 730).

Besides the items about the perceived course difficulty, workload, pace, expected grade, and student's prior interest in the course, part three in SEEQ also collects demographic information about the student and the teacher as well as background information about the course. The last part of the form, part four, contains two open-ended questions, which ask students to provide comments about the strengths and areas for improvement for the teacher being rated, and space for additional comments and/or clarifications of any responses to the scale items.

While the research leading to the development of SEEQ was carried out in the North American context, numerous studies have been carried out to explore the applicability of SEEQ in different countries, including Australia (Hayton, 1983; Marsh, 1981; Marsh & Roche, 1993), Spain (Marsh, Touron, & Wheeler, 1985), Papua New Guinea (Clarkson, 1984), USA (Marsh & Hocevar, 1984, 1991a), India (Watkins & Thomas, 1991), New Zealand (Watkins, Marsh, & Young, 1987), Hong Kong (Watkins, 1992), Nepal (Watkins & Regmi, 1992), Nigeria (Watkins & Akande, 1992), the Philippines (Watkins & Gerong, 1992), China (Lin, Watkins, & Meng, 1995; Marsh, Hau, Chung, & Siu, 1998), Taiwan (Lin, Watkins, & Meng, 1994), UK (Coffey & Gibbs, 2001), and Jamaica (Penny, 2004). There is strong evidence in these studies for the applicability of the

SEEQ factors outside the North American context in which the instrument was developed (Watkins, 1994; Marsh, 2007). In summary, SEEQ is the most widely used instrument in published work (Richardson, 2005), "with a robust factor structure, excellent reliability and reasonable validity" (Coffey & Gibbs, 2001: 89).

For ease of interpretation of data and to facilitate comparability of findings between different studies in the field, the use of standardised rating forms, as opposed to non-standardised forms which may be confounded by reliability and validity problems, has long been advocated in SET research (L'Hommedieu, Menges and Brinko, 1990; Richardson, 2005). Research on SEEQ in America and Australia suggests that it is possible, and indeed feasible, to construct a SET instrument that has a very wide range of applicability and which can be used to make meaningful comparisons across a wide variety of academic disciplines and higher education institutions (Richardson, 2005). Richardson also adds that "such a questionnaire should be motivated by research evidence about teaching, learning and assessment in higher education and that it should be assessed as a research tool" (p.404).

Permission to use SEEQ for research purposes was obtained from its developer Professor Herbert March via personal correspondence in March 2008.

5.4 Research Design

The purpose and type of the investigation and the research questions formulated after the exploratory study, along with research ethics, guided the sampling strategies and data collection procedures in this study.

5.4.1 Sampling

Access to students and lecturers was negotiated with the Directors of English Language Centres in the selected colleges. Data on participants' perceptions, opinions and student ratings were collected from a representative sample of program administrators, lecturers and students in the GFPs in six Colleges of Technology in Oman during term three of the academic year 2007-2008.

A mixture of cluster sampling and quota sampling was used to select the students' sample in the study. First of all, the existing levels of English proficiency in each GFP Program were identified. Then a sample of groups from each level were randomly selected to complete the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire, in proportion with the total number of groups in that level. From the six colleges, 968 students completed this questionnaire representing three different GFP levels (see Table 5.2).

Table 5.2: Distribution of Student Respondents to the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* Questionnaire

LEVEL	GENDER		Total
	Male	Female	
Elementary	169	28	197
Intermediate	239	145	384
Advanced	209	178	387
Total	**617**	**351**	**968**

Simple random sampling was used to sample lecturers. The *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire was distributed to all the lecturers available at the colleges during the time of administration. A total of 248

completed questionnaires were returned, around 60% of the total population of lecturers in the GFPs in the six colleges. A large sample was used to ensure that lecturers from different ethnic and linguistic backgrounds are represented (see Table 5.3).

Unlike the general population of students and lecturers, the population of Program administrators is very small. Each GFP is managed by a team of three administrators, namely: the Director of the Language Centre, the Head of Section of Curriculum & Teaching Methods, and the Head of Section of English Language Programs. All were invited to complete the teacher's version of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire. Based on the findings of the exploratory qualitative study, the differences in perceptions found between teachers and administrators were minor compared to the differences observed between both of these two groups on one side and students on the other. For this reason, and also because GFP administrators are technically teachers with extra administrative responsibilities, it was decided to include them with the teachers' sample.

Table 5.3: Distribution of Teacher Respondents to the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* Questionnaire

	Gender	Teacher's Mother Tongue											
Teacher's Ethnicity		Arabic	English	Urdu	Hindi	Malayalam	Tamil	Tagalog	Other South Asian or Southwest Asian language	African	European	Other	Total
Omani Arab	Male	9							1				10
	Female	9			1				1				11
Non-Omani Arab	Male	22											22
	Female	8											8
South Asian or Southwest Asian	Male		1	7	7	13	10		5				43
	Female		7	7	7	13	8		7				49
European	Male		7								2		9
	Female		6								1		7
African	Male		4							5			9
	Female		5								1		6
North American	Male	1	14									1	16
	Female		23									1	24
Southeast Asian	Male							10					10
	Female							7					7
Other	Male												
	Female		4		1								5
	Total	49	71	14	16	26	18	17	14	5	4	2	236
												Cases with missing values	12
												Total	**248**

Then, SEEQ was administered to the same groups of students that completed the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire two weeks earlier. For each group of students, one of the English courses they were taking that term was randomly selected for rating with SEEQ (See Table 5.4).

Table 5.4: SEEQ Respondents Cross-tabulated by Course, Level, and Gender

Course	Level	Gender		Total
		Male	Female	
Core Course	Elementary	15	5	20
	Intermediate	105	59	164
	Advanced	30	30	60
Listening & Speaking Skills	Elementary	62	2	64
	Intermediate	35	28	63
	Advanced	69	61	130
Reading Skills	Elementary	49	16	65
	Intermediate	50	12	62
	Advanced	47	36	83
Writing Skills	Elementary	31	5	36
	Intermediate	34	39	73
	Advanced	50	52	102
	Total	577	345	**922**

In this round, 922 students completed SEEQ. After the lecturers were randomly selected by the researcher, permission for the rating exercise to take place was sought from each individual lecturer.

Table 5.5: Sample Summary for the Main Study

Sub-sample	Instrument	Total No. of Respondents	The Perceptions Questionnaire ONLY	SEEQ ONLY	Both
Students	*Perceptions of Good College Teaching & Students' Evaluation of Teachers*	**968**	**94**	-	

	Students' Evaluation of Educational Quality ***(SEEQ)***	**922**	-	**48**	**874**
Teachers	*Perceptions of Good College Teaching & Students' Evaluation of Teachers*	**248**			
	Total	**2138**			

As can be seen from Table 5.5, 94 students out of the 968 who completed the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* were absent when the SEEQ was administered and, therefore, did not participate in the SET rating. One the other hand, of the 922 students who completed the SEEQ, 48 students were absent when the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* was administered. Therefore, the number of students who completed both questionnaires is 874.

5.4.2 Data Collection

Using the teachers' and students' versions of the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire developed by the researcher and the standardised SEEQ rating questionnaire explained above, the actual data collection for this phase of the study started on 2nd April 2008 and continued until the 30th May 2008.

This design was considered suitable for the study for a number of reasons. It is only during term time that program administrators, lecturers and students are available to

participate in the study. Also, generalisation of findings to a wider population demands a big, representative sample with reliable and valid instruments that fit the aims and questions of the study and nature of the enquiry. Equally important, economical designs, such as Likert scale surveys, were considered more suitable for the purpose.

The researcher avoided conducting the study in terms one and two because CTs usually receive their student intakes at the beginning of these two terms- September and January. From the researcher's own experience, during these two terms, some of the GFPs usually face shortfall in lecturers and physical resources and, consequently, resort to temporary measures such as merging groups, mobilising lecturers across levels and courses, or modifying delivery plans to cover the shortage in lecturers and resources. The researcher's judgment was that such potential instability in the Program during these times would jeopardise the data collection plan and the quality of data collected. In extremely busy times like these, approaching the Directors of GFPs for assistance with research could add to the problems they may already have, especially when the questionnaires are administered in two rounds and the sample of lecturers and students involved is big. This in turn could lower the level of their cooperation with the researcher.

Questionnaires were administered in two stages. In the first stage, which was executed in weeks 3 & 4 of the summer semester, the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire was administered to all lecturers available in the language centres at the time of administration and to 46 GFP classes in six CTs. To ensure consistency in administration procedures and instructions, the

researcher personally supervised the distribution and collection of questionnaires from students and lecturers. Students were given 20 minutes to complete the questionnaire during class time. At the end of this round, 980 student questionnaires and 253 teacher questionnaires were collected.

As far as questionnaire administration is concerned, Brown (2001) makes a distinction between two types of questionnaires: self-administered questionnaires and group-administered questionnaires. Brown argues that group-administered questionnaires, which are administered to groups of respondents all at one time and place like the questionnaires used in this phase of the study, have a number of advantages over self-administered questionnaires. Because group-administered questionnaires are usually administered to captive participants, like students in a class, the return rate tends to be high, as respondents feel obliged to respond an also because it becomes easy to track down absentees and ask them to fill in the questionnaire. Group administration also enables the researcher to be present to explain any ambiguities or make any clarifications that may be required. In addition, administering the questionnaire in groups enables the researcher to know and control the conditions under which the questionnaire was filled out.

Stage two took place in weeks 5 and 6 of the same semester. During class time, every group of the 46 groups that completed the perceptions questionnaire two weeks earlier was asked to anonymously rate one of their TESOL lecturers teaching them in that semester. As mentioned earlier, the teachers who were rated were selected randomly by the researcher and their consent for the ratings to be collected was obtained before hand.

Students in each group were given 20 minutes to complete the rating questionnaires. Before the rating forms were distributed, lecturers were kindly requested to leave the room. At the end of this round, 931 SEEQ questionnaires were collected.

Before the administration of both questionnaires, students were assured that the data they provided in the background information section of both questionnaires was only for data organisation purposes and would not be used to identify the student in the research report. Students were also encouraged to give frank and honest ratings in the SEEQ and were promised that their identities, ratings and comments as individual students would not be disclosed to their lecturers or colleges and would be used for research purposes only.

5.5 Data Analysis

Following the researcher's return to the UK in early June 2008, work on data organisation and entry started. Bearing in mind the size of the sample and the total number of questionnaires used (N= 2138), the process of data entry required a lot of time and patience. Prior to the main data analysis, a preliminary data screening and analysis was conducted to evaluate the completeness and suitability of data for analysis.

5.5.1 Data Screening and Preliminary Analysis

Firstly, all the questionnaires were checked for faults, such as missing pages, unusually high frequency of user-missing values, or blank background information section. Following Overall & Marsh (1979), questionnaires which were less than 75% complete were excluded from the analysis. In the rest of the questionnaires, variables with missing

values were deleted pair-wise or analysis-by-analysis. Due to the group administration technique used in this study and the researcher's close and personal supervision of the questionnaire distribution and collection processes, the number of questionnaires which were less than 75% complete was very small. In the case of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire, only 12 students' questionnaires (out of 980 originally collected) and 5 teachers' questionnaires (out of the 253 received from the teachers) were excluded from the analysis due to low completion rates.

Secondly, every questionnaire was given a case number that differentiates it from the rest of the sample. For student participants it was not possible to use their ID number as a case number, because ID numbers were duplicated and not unique, since the vast majority of students completed both questionnaires, the perceptions questionnaire and the SEEQ. Thirdly, every item on both questionnaires was coded and given a variable name. After that, data from the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire was entered onto datasets using Statistical Package for the Social Sciences (SPSS) for Windows, version 15.0, for both students and lecturers.

A similar process was followed to screen, organise and enter the quantitative data from SEEQ questionnaires onto SPSS: checking for faults and missing values, numbering of questionnaires, coding of variables, and finally creating an SPSS dataset. Besides screening for missing values, the data was also checked for the shape of distribution and the presence of outliers that could seriously distort the results. Only nine out of the 931

SEEQ questionnaires originally distributed were eliminated from the analysis due to low completion rate. Variables with missing values were deleted pair-wise or analysis-by-analysis. Bearing in mind the research questions and objectives of the investigation in hand with regard to SEEQ, the two open-ended questions in this instrument were not included in the data analysis, as including them serves neither the factor analysis nor the reliability analysis, which are the focus of data analysis in most of SEEQ's applicability studies discussed in Chapter 4.

After the completion of data entry for the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire, and the initial data screening, a preliminary data analysis was carried out to determine the reliability coefficient of its two scales using SPSS. The results were as follows:

Table 5.6: Preliminary Analysis of the Reliability of the Scales Used in the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* Questionnaire

Version	**Section 1** *Perceptions of the characteristics of good college teachers* scale	**Section 2** *Perceptions of SET* (Mean of 3 sub-scales)
Teachers	.95	.76
Students	.89	.74

As can be seen in Table 5.6, the alpha values for section one scale seem to be of a good standard. The mean reliability coefficients for the 3 sub-scales of section 2, on the other hand, are moderate. These values are, nonetheless, higher than those found in the pilot version of the questionnaire reported earlier under section 5.3.2.1.

As pointed out earlier, SEEQ scales were not subjected to a preliminary reliability analysis because two of the main research questions were exclusively designed to investigate the reliability and factor structure of the SEEQ in Oman. However, prior to the factor analysis, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Barlette's test of sphericity were calculated to determine the degree of variance among the different items and the overall significance of correlations between them. These two tests are used to determine the suitability of SEEQ scales for factor analysis (Penny, 2004). The results of this analysis indicated that it was appropriate to carry out the factor analysis on the SEEQ ratings collected from the students in Oman. The results of the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Barlette's test of sphericity are given in Chapter 8.

5.5.2 Primary Analysis

Data from sections 1 and 2 in both versions of the *Perceptions of Good College Teaching & Students' Evaluation of Teachers* questionnaire were used to answer the main research questions number 1, 2, 3, and 4. Consistent with the methods used by other researchers in the field, the item mean is used to identify the ranking of the different characteristics of effective college teaching embedded in each item. Spearman's rho correlation coefficient was used to determine the power of correlation between teachers' and students' rank-ordering of the characteritics of effective teaching. Students' and lecturers' rankings were also compared using independent sample *t*-tests to identify the significance of matches/mismatches between the rankings or perceptions of the two groups.

Pearson's chi-square was used to test for differences in perceptions of effective teaching between male and female students. Kruskal-Wallis tests and Mann-Whitney U tests were also used to test for differences based on ethnicity or mother tongue between the perceptions of different groups in the teacher population.

Consistent with the methods used in SEEQ applicability research, an exploratory factor analysis and an inter-rater reliability analysis were performed on group ratings to identify SEEQ factor structure and the reliability of the instrument in the Omani context (research questions 5&6). According to Marsh and Hocevar (1991b), SEEQ factor analysis serves different purposes and tests whether: (a) students are capable of differentiating among the multi dimensions of teaching, (b) the factors emerging from the data confirm the ones the instrument is designed to measure, and (c) the same factors are being identified consistently across varying settings and disciplines. In addition to all of these functions, factor scores can also be used to summarise and report the results of students' ratings.

The factor analysis carried out on SEEQ in this study applied similar procedures used by Marsh and Hocevar (1991b). The number of factors to be extracted was limited to eight- the number of SEEQ factors excluding the Workload/Difficulty scale. The factors were extracted using the principal axis factoring extraction method on SPSS, version 15.0. This was followed by Varimax rotation to reach a 'simple structure', "minimising the number of items that load highly on a factor" (Penny, 2004: 183-184).

Following the factor analysis of SEEQ, there was a reliability analysis. Consistent with the objectives of this study and the recommendations of the research literature in the subject, the inter-rater reliability, or the level of agreement among students of the same class rating the same aspect of effective teaching (Marsh & Dunkin, 1992; Cashin, 1995) was estimated. The present reliability analysis applied the same procedures used by Marsh and Roche (1993). Inter-rater reliability was estimated with *MLwiN*, version 2.17 (Rasbash, Charlton, Browne, Healy and Cameron, 2010), from the class-average response using the intraclass correlation coefficient obtained with a one-way analysis of variance (Penny, 2004). The analysis is carried out for the total scale and for factor scales, for the total sample and then for different GFP levels and courses.

Spearman's rho correlation coefficient, Kruskal-Wallis tests, and Mann-Whitney U tests were also used to test for relationships between student, teacher, and course background variables and teacher's overall rating or to test for differences between groups based on these background variables.

5.6 Research Ethics

Codes of educational research ethics discussed in the literature (e.g. Coomber, 2002; Pring, 2001; Small, 2001; Wiles, Charles, Crow & Heath, 2006) were carefully observed by the researcher in planning the study, in designing and executing it, and in analysing and reporting its results. Ethical issues pertaining to the individuals participating in this study, the institutions sampled, the general public, and academic honesty in the field and after were carefully studied and adhered to.

During the planning phase, approval of the research ethics by the School of Education Ethics Committee, University of Durham, was sought and obtained. In addition, the researcher took two online courses on ethics in social research and attended a further workshop on the same, all offered by the University of Durham, prior to engaging in field work. All these steps were taken by the researcher in recognition of his responsibility towards the participants and all the institutions involved in the investigation.

In designing and carrying out the research, measures were taken to ensure that participants were protected from harm and respected. Access to the participants and consent to carry out the research were sought and obtained from the Directors of the GFPs or the Deans of the colleges involved. In addition, informed consent from the participants was obtained upon the administration of the questionnaires. The covering letters of the questionnaires explained the purpose of the research and reassured the participants of the confidentiality of their views and the data they provided. In these letters, the rights of the respondents to participate or withdraw from the study at any point and without prejudice were clearly stated (see Appendices 3, 4, 5, 7, 8). In addition to the letters, these rights and reassurances were reiterated verbally by the researcher in front of every group of students before they completed the questionnaires. As pointed out earlier under the data collection section, time for questionnaire administration was chosen carefully and negotiated with the management of the GFPs to minimise disruption to classes. Where students' ratings of a course were collected, the concerned teacher's consent was obtained first. To ensure the highest level of confidentiality in students' ratings and to encourage students to rate their teachers with confidence and

honesty, it was agreed with the volunteering teachers that the rating forms would be distributed to their students by the researcher only after the concerned teacher has left the classroom. As far as the use of the standardised rating instrument is concerned, permission to use the SEEQ was obtained from the developer of the questionnaire, Professor Herbart Marsh, before the commencement of the field work.

In analysing and reporting the results, a number of steps were taken to protect the confidentiality of the participants and the openness and accessibility of the study to the participants and the other researchers in the field. No names of participants or participating institutions were revealed in the research report. In addition, all the data collection procedures and data analyses processes were explained in detail for verification or replication purposes.

CHAPTER SIX

RESULTS AND DISCUSSIONS: STUDENTS' AND TEACHERS' PERCEPTIONS OF THE IMPORTANCE OF VARIOUS CHARACTERISTICS OF EFFECTIVE COLLEGE TEACHING: MATCHED OR MISMATCHED PRIORITIES?

6.0 Introduction

In Chapters 1-4, an introduction to the study and a review of the relevant literature were presented. In Chapter 5, the research methodology and design for both stages of the study have been discussed. Chapters 6-8 report on the results of data analysis and provide discussions of the findings. The presentation of data and discussion of the findings in these three chapters is based on the themes generated from the research objectives and research questions identified in Chapter 1. For ease of presentation, each of these themes will be allocated a separate chapter in which the results of data analysis and discussions of the findings to the relevant research questions will be presented in an integrated manner. It is hoped that this will enable the reader to easily establish a link between the research objectives and questions, the data gathered, and the discussions of the findings, and will assist in understanding the connection between these findings and the literature review.

Chapter 6 presents and discusses the findings about the match and mismatch between students' and lecturers' perceptions of the importance of various characteristics of effective college teaching in the six Omani colleges of technology surveyed in this

study. Following this, in chapter 7, the data about students' and lecturers' perceptions of the factors hypothesised to bias SET, the utility of students' evaluation of college teaching, and the role of students as evaluators of teaching is presented and discussed. Chapter 8 reports on the analysis of the data collected with SEEQ from the sampled colleges and discusses the factor structure underlying the SEEQ ratings in Oman. Included in Chapter 8 also is a discussion of the findings about the reliability of SEEQ in the Omani context and the potential effect of various course, teacher, and student characteristics on students' ratings.

To this end, this chapter presents and discusses the findings to the following research questions:

- **Research Question 1:** To what extent do students' and teachers' perceptions of the importance of various characteristics of effective college teaching match or mismatch?

- **Research Question 2:** To what extent does students' gender have an effect on their perceptions of the importance of various characteristics of effective college teaching?

- **Research Question 3:** To what extent do mediating factors such as teachers' ethnic background and mother tongue have and effect on their perceptions of the importance of various characteristics of effective college teaching?

As mentioned earlier in chapter five, the data used to answer these questions were collected using the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire (Appendices 7 & 8) that was developed following a comprehensive literature review and a qualitative exploratory investigation whose participants came from four out of the six colleges that participated in the main study.

6.1 Research question 1: To what extent do students' and teachers' perceptions of the importance of various characteristics of effective college teaching match or mismatch?

As explained in Chapter 5, respondents' perceptions of the ranking in importance of various dimensions of effective college teaching were identified using Likert-type ranking scales (see Section One in Appendices 7 & 8). A total of 38 characteristics or dimensions (Table 6.1) were identified and included in section one of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire. This section of the questionnaire measured lecturers' and students' rank-ordering of the importance of each of the characteristics listed on a 5-point scale – "(1) not at all important" to "(5) extremely important".

The Statistical Package for Social Sciences (SPSS), version 15, and Microsoft Excel 2003 were used to analyse the data resulting from this section of the questionnaire. A number of procedures were followed to identify the differential perceptions of lecturers and students of the importance of the 38 characteristics of effective college teaching and the degree to which these perceptions matched or mismatched. These ranged from direct comparisons between lecturers' and students' rankings of the importance of these dimensions of good teaching drawn from descriptive statistics to other statistical techniques aimed at assessing the strength of correlation between lecturers' and students' ranking and identifying the significance level of the observed differences between them.

In line with the data analysis techniques used in similar studies, the differential ranking of the importance of each item was derived from the item mean. Item means were then ranked for students and lecturers and compared and correlated using Spearman's rho correlation coefficient. Independent Sample *t*-test was used to identify statistically significant differences in means between students' and lecturers' perceived ranking of the importance of each characteristic of effective teaching.

Table 6.1: Thirty-eight Characteristics of Effective College Teaching and Their Abbreviations

	Characteristics of Effective College Teaching	Abbreviation
1.	Encouraging students to participate in classroom activities and discussions	Encourage participation
2.	Inviting students to share their ideas and knowledge	Invite ideas
3.	Demonstrating good skills in classroom management	Classroom management
4.	Demonstrating very good use of student-centred approaches	Student-centred
5.	Encouraging students to ask questions and ensuring that answers given to students are meaningful	Encourage questions
6.	Encouraging students to express their own ideas and/or question the lecturer	Encourage expression
7.	Presenting the background or origin of ideas/concepts developed in class	Build on knowledge
8.	Demonstrating a high level of expressiveness and giving clear explanations	Expressiveness
9.	Giving lectures/tutorials in a style and pace that facilitate note-taking	Pace
10.	Showing flexibility and diversity in teaching style	Flexibility and diversity
11.	Preparing good course materials and carefully explaining them to students	Preparation
12.	Presenting points of view other than lecturer's own when appropriate	Different view points
13.	Using lively presentation styles which hold students' interest during class	Lively presentation
14.	Making proper use of instructional media, teaching aids, and multi-media labs	Use teaching aids
15.	Enhancing presentation with the use of humour	Humour
16.	Showing strong enthusiasm for the subject	Enthusiasm
17.	Making the course intellectually challenging and stimulating	Challenge
18.	Being able to stimulate the interest of the students in the subject	Stimulate interest
19.	Showing dedication to teaching	Dedication
20.	Being dynamic and energetic in conducting the class	Dynamic &

		energetic
21.	Using appropriate and fair methods of evaluating student work	Fair evaluation
22.	Giving valuable feedback on assessments/graded material	Feedback
23.	Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject	Valuable homework
24.	Demonstrating full compliance with the announced objectives of the course	Comply with objectives
25.	Giving assignments/graded materials which test class content as emphasised by the lecturer	Test content
26.	Having relevant Academic qualifications in teaching English as a second/foreign language	Qualifications
27.	Having sufficient formal teacher training in teaching English as a second/foreign language	Well-trained
28.	Keeping abreast of the latest developments in the field or subject	Up-to-date
29.	Having relevant and sufficient experience in teaching English as a second/foreign language	Experience
30.	Having full command of the subject matter	Subject mastery
31.	Making students feel welcome in seeking help/advice in or outside of class	Welcome help requests
32.	Having a genuine interest in individual students	Genuine interest
33.	Showing respect for all students	Respect
34.	Being available to students for advice and support in or after class	Available for advice
35.	Showing sensitivity to the culture of the organisation and society at large	Sensitivity to culture
36.	Being friendly toward individual students	Friendly
37.	Having native-like intonation and stress	Stress & intonation
38.	Being a native speaker of the target language	Native speaker

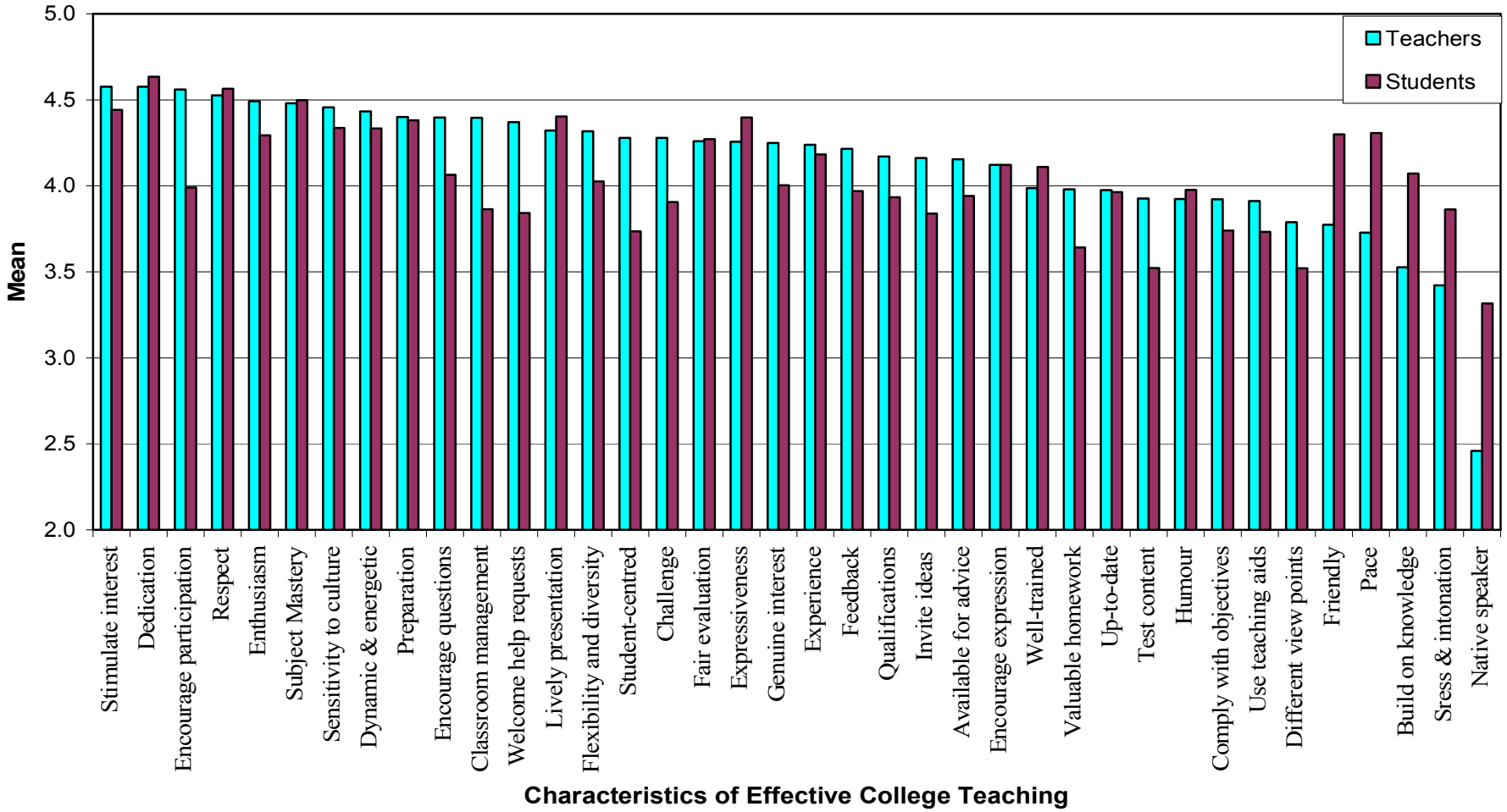
The results and discussions of the findings for this question will start with an examination of the overall correlation between lecturers' and students' perceptions of the degree of importance they attach to each of the 38 characteristics of effective college teaching. Following this will be a detailed presentation and discussion of the differential levels of importance of each specific characteristics of effective teaching as perceived by both students and lecturers. These will be compared and contrasted to identify the matches and mismatches in ranking between the two sample groups.

6.1.1 Overall Correlations between Teachers' and Students' Ranking of the Importance of the Characteristics of Effective College Teaching

Separate group means were calculated for the lecturers' and students' ratings of each of the characteristics listed above. Figure 6.1 shows all the characteristics ranked by magnitude of rated importance for both sample groups. Following Feldman's (1988) approach, after determining the differential importance of the various characteristics of effective teaching for both the lecturers and the students, the rankings for the two groups were correlated to assess the degree of overall agreement/ disagreement between them. The relationship between students' (N= 968) and teachers' (N= 248) rank-ordering was investigated using Spearman's rho correlation coefficient, which is considered more suitable for ranked or ordinal level rating Likert scales (Pallant, 2007).

A moderate but statistically significant overall correlation, *rho*= .56, $p<.001$, was found between teachers' and students' differential ranking of the 38 characteristics of effective college lecturers. This overall correlation strength is lower than the .71 reported by Feldman (1988). When examining the results from the type of sample or specialisation of the respondents' point of view, the correlation coefficient obtained from the sample in the present study (TESOL foundation program) is also much lower than the correlation coefficient for social sciences ($r = +.88$), the humanities ($r = +.85$), and engineering ($r = +.80$) reported by Marques *et al.* (1979).

Figure 6.1: Mean Teachers' and Students' Perceptions of the Importance of 38 Characteristics of Effective Teaching



However, the correlation power between students' and lecturers' ranking found in this study is stronger than what Baum and Brown (1980), Stevens (1978), Stevens and Marquette (1979), and Wotruba and Wright (1975) reported in their studies involving students and staff in business schools, which was an average of r=.26. One of the reasons why the correlation coefficient found in this study was relatively lower than that previously found could be attributed to the wide diversity in the lecturers' population sampled in the present study. Unlike student participants in this study, who are almost exclusively Omani, lecturers in the GFPs come from diverse backgrounds, contexts, and nationalities and many of them had never taught Omani students before taking their teaching posts at the sampled institutions.

However, as highlighted in the literature review in chapter 3, assessing the overall correlations between students' and faculty's rank-ordering of the characteristics of effective teaching, although important, is not sufficient to explore the matches and mismatches between the two. Identifying the weights each side attaches to various specific dimensions of effective teaching requires a different level of analysis in which a detailed examination of the differential ranking of dimensions by students and their teachers is carried out. This is the theme of the following section.

6.1.2 Teachers' and Students' Differential Ranking of the Importance of the Characteristics of Effective College Teaching

This section will present and discuss the differential ranking of each of the specific characteristics of effective teaching as perceived by lecturers and students. As evident in Figure 6.1, and in agreement with the findings of many other studies (e.g. Feldman, 1988; Fisher, Alder, and Avasalu, 1998; Raymond, 2008) there are many similarities between the rankings of the two groups, but they are not identical.

6.1.2.1 The Mean as a Point of Comparison

When taking the group mean as a point of comparison (Figure 6.1 & Table 6.2) between students and teachers, it can be seen that the lecturers placed greater importance on 24 of the 38 characteristics. These are: stimulating students' interest in the course (*Mean*= 4.58), encouraging students' participation (*Mean*= 4.56), showing enthusiasm for the course (*Mean*= 4.49), showing sensitivity to the culture of the organisation and the society (*Mean*= 4.46), being dynamic and energetic in conducting the class (*Mean*=4.43), preparing good course materials (*Mean*= 4.40), encouraging students to ask questions (*Mean*= 4.40), demonstrating good skills in classroom management (*Mean*= 4.40), welcoming requests for help from students (*Mean*=4.37), showing flexibility and diversity in teaching style (*Mean*=4.32), using student-centred approaches (*Mean*=4.28), making the course intellectually challenging (*Mean*=4.28), showing genuine interest in individual students (*Mean*=4.25), having relevant and sufficient experience in teaching the subject (*Mean*= 4.24), giving valuable feedback (*Mean*= 4.22), having relevant academic qualifications in the subject (*Mean*=4.17), inviting students to share their ideas (*Mean*=4.16), being available for advice and support (*Mean*=4.15), assigning valuable homework (Mean=3.98), keeping abreast of the latest developments in the field (*Mean*=3.98), giving assignments which test class content as emphasised by the teacher (*Mean*=3.93), complying with the announced objectives of the course (*Mean*=3.92), Using teaching aids *(Mean*=3.91), and presenting points of view other than teacher's own (*Mean*=3.79).

On the other hand, students ranked 13 characteristics higher than the lecturers did. These are: dedication to teaching (*Mean*= 4.63), respect (*Mean*=4.56), mastery of the

subject matter (*Mean*=4.50), giving lively presentations (*Mean*=4.40), fair evaluation (*Mean*=4.27), expressiveness (*Mean*=4.40), having sufficient formal teacher training (*Mean*= 4.11), enhancing presentation with humour (*Mean*=3.98), being friendly (*Mean*=4.30), giving lectures in a pace which facilitates note-taking (*Mean*=4.31), building on previous knowledge (*Mean*=4.07), having native-like intonation and stress (*Mean*=3.86), and being a native speaker of the target language (*Mean*=3.32). On one criterion, *encouraging expressio*n, the means of the two groups were exactly the same (*Mean*=4.12).

6.1.2.2 The Rank as a Point of Comparison

When examining the matches and mismatches between the students and the teachers in their perceptions of the importance of the 38 characteristics of effective teaching using the rank as a point of comparison, some of the mismatches between the two groups appear more striking. Seven rank differences are particularly large. These can be summarised as follows:

- While lecture pace was ranked 35 by the teachers, it was ranked 10 by the students. This is a rank difference of 25.
- Being friendly was ranked 34 by the teachers, but only 11 by the students- a rank difference of 23.
- Building on previous knowledge was ranked 36 by the teachers, but ranked 17 by the students, which is a difference of 19 ranks.
- There was also a difference of 19 ranks in rating the importance of demonstrating good skills in classroom management between the teachers (rank 9) and students (rank 28).
- While welcoming requests for help was ranked 12 by the teachers, it was ranked 30 by the students. This is a rank difference of 18.

- Encouraging students to participate in classroom activities was ranked 3 by the teachers, but only 21 by the students- a difference of 18 ranks.
- There was a difference of 17 ranks between teachers' (rank 15) and students' (rank 32) rank-ordering of the importance of using student-centred approaches in teaching.

However, there were two perfect matches and 15 other close matches between the teachers and the students in their rank-ordering of the importance of the criteria of effective teaching. Both teachers and students assigned the top rank to dedication to teaching. Both groups also saved the last rank, 38, to being a native speaker of the target language. The fifteen close matches -with rank differences of only 1-5 ranks- were found for the following traits: stimulating students' interest, respect, mastery of the subject matter, sensitivity to the culture of the organisation, being dynamic and energetic, good preparation, fair evaluation, having genuine interest in students, giving valuable feedback, having proper academic qualifications, being available for advice, keeping abreast of the latest developments in the field, complying with the course objectives, using teaching aids, and presenting points of view other than teacher's own.

The finding about lecturers opting to attach higher importance to a bigger number of criteria in judging effective teaching compared to students is not unique to this study. It was also reported by Fisher *et al.* (1998), where 19 out of the 22 characteristics of good teaching included in their study were rated higher by the lecturers.

6.1.2.3 Testing for Significant Differences

While the analysis given above gives an indication of the matches and mismatches between the teachers and the students in their valuation of the different dimensions of effective teaching based on the group mean and ranks, it does not provide evidence of the statistical significance of the differences between the two groups. In order to assess the significance of the observed mismatches in importance rankings, an independent Sample *t* Test was used (Table 6.2). Because the number of dependent variables to be analysed was relatively high and, as a result, the chance to make a Type I error was increased (Field, 2009; Fisher *et al.*, 1998; Pallant, 2007), there was a need to protect the alpha level. The alpha level was adjusted to $p = .001$ instead of the traditional .05 (.05/38= .001) using the Bonferroni adjustment technique (Pallant, 2007; Tabachnick and Fidell, 2007). In addition, the *t*-value and its associated significance level of differences between group means for each dependent variable were verified against the results of Levene's test of equality of variances accompanying the independent t-test. Where Levene's test indicated a violation of the assumption of equal variances for any of the dependent variables, the alternative *t*-value and its associated *p* value provided by the test were used.

Table 6.2: Comparisons of Teachers' and Students' Importance Rankings of the 38 Characteristics of Effective College Teaching

	Teachers						Students						M	R	*t*	Sig.*
Characteristic	**N**	**Min**	**Max**	**M**	**SD**	**R**	**N**	**Min**	**Max**	**M**	**SD**	**R**	**Diff.**	**Diff.**		**(2-tailed)**
Stimulate interest	246	2	5	4.58	0.63	1	956	1	5	4.44	0.81	4	0.14	-3	2.81	.005
Dedication	248	1	5	4.58	0.66	1	961	1	5	4.63	0.69	1	-0.05	0	-1.18	.239
Encourage participation	248	3	5	4.56	0.58	3	965	1	5	3.99	1.01	21	0.57	-18	11.58	**.000**
Respect	247	2	5	4.53	0.65	4	954	1	5	4.56	0.73	2	-0.03	2	-.76	.448
Enthusiasm	248	2	5	4.49	0.64	5	966	1	5	4.29	0.77	12	0.20	-7	4.16	**.000**
Mastery of subject	248	1	5	4.48	0.69	6	967	1	5	4.50	0.77	3	-0.02	3	-.35	.728
Sensitivity to culture	245	2	5	4.46	0.64	7	965	1	5	4.34	0.89	8	0.12	-1	2.41	.016
Dynamic & energetic	247	2	5	4.43	0.62	8	962	1	5	4.33	0.82	9	0.10	-1	2.10	.037
Preparation	247	2	5	4.40	0.71	9	964	1	5	4.38	0.76	7	0.02	2	.36	.722
Encourage questions	247	2	5	4.40	0.63	9	958	1	5	4.06	0.88	18	0.34	-9	6.80	**.000**
Classroom management	245	2	5	4.40	0.62	9	962	1	5	3.86	1.05	28	0.54	-19	10.24	**.000**
Welcome help requests	248	2	5	4.37	0.70	12	960	1	5	3.84	1.00	30	0.53	-18	9.62	**.000**
Lively presentation	246	2	5	4.32	0.70	13	963	1	5	4.40	0.86	5	-0.08	8	-1.58	.115
Flexibility and diversity	246	2	5	4.32	0.70	13	952	1	5	4.03	0.94	19	0.29	-6	5.37	**.000**
Student-centred	247	2	5	4.28	0.74	15	966	1	5	3.74	1.03	32	0.54	-17	9.41	**.000**
Challenge	247	2	5	4.28	0.72	15	964	1	5	3.91	0.94	27	0.37	-12	6.81	**.000**
Fair evaluation	247	2	5	4.26	0.71	17	965	1	5	4.27	0.89	13	-0.01	4	-.23	.817
Expressiveness	246	2	5	4.26	0.67	17	963	1	5	4.40	0.83	5	-0.14	12	-2.82	.005
Genuine interest	248	2	5	4.25	0.76	19	964	1	5	4.00	1.02	20	0.25	-1	4.21	**.000**
Experience	247	1	5	4.24	0.76	20	966	1	5	4.18	0.94	14	0.06	6	.98	.329
Feedback	246	1	5	4.22	0.70	21	963	1	5	3.97	0.97	23	0.25	-2	4.51	**.000**
Qualifications	247	1	5	4.17	0.90	22	958	1	5	3.93	1.08	26	0.24	-4	3.52	**.000**
Invite ideas	247	1	5	4.16	0.73	23	963	1	5	3.84	1.04	30	0.32	-7	5.64	**.000**

Available for advice	247	2	5	4.15	0.72	24		948	1	5	3.94	0.99	25		0.21	-1	3.80	**.000**
Encourage expression	246	2	5	4.12	0.75	25		967	1	5	4.12	0.93	15		0.00	10	-.00	.999
Well-trained	247	1	5	3.99	0.88	26		964	1	5	4.11	1.00	16		-0.12	10	-1.89	.059
Valuable homework	247	2	5	3.98	0.74	27		965	1	5	3.64	1.00	35		0.34	-8	5.92	**.000**
Up-to-date	246	1	5	3.98	0.86	27		955	1	5	3.96	0.93	24		0.02	3	.19	.851
Test content	247	1	5	3.93	0.77	29		966	1	5	3.52	1.10	36		0.41	-7	6.70	**.000**
Humour	247	1	5	3.92	0.85	30		935	1	5	3.98	1.05	22		-0.06	8	-.85	.394
Comply with objectives	245	1	5	3.92	0.84	30		963	1	5	3.74	1.03	32		0.18	-2	2.89	.004
Use teaching aids	247	1	5	3.91	0.82	32		966	1	5	3.73	1.01	34		0.18	-2	2.91	.004
Different view points	247	2	5	3.79	0.78	33		960	1	5	3.52	1.05	36		0.27	-3	4.48	**.000**
Friendly	243	1	5	3.77	1.05	34		966	1	5	4.30	0.89	11		-0.53	23	-7.18	**.000**
Pace	246	1	5	3.73	0.84	35		962	1	5	4.31	0.88	10		-0.58	25	-9.33	**.000**
Build on knowledge	245	1	5	3.53	0.93	36		965	1	5	4.07	0.93	17		-0.54	19	-8.17	**.000**
Stress & intonation	242	1	5	3.42	1.05	37		964	1	5	3.86	1.06	28		-0.44	9	-5.80	**.000**
Native speaker	246	1	5	2.46	1.27	38		966	1	5	3.32	1.44	38		-0.86	0	-9.16	**.000**

*** *p*<.001**

N= Number of cases**; Min=** Minimum value**; Max=** Maximum value**; M**= Mean; **SD**= Standard Deviation; **R**= Rank; **M Diff.**= Difference in means; **R Diff.**= Rank Difference.

Note: The **mean difference** for each dimension was obtained by subtracting students' mean from teachers' mean. A positive value indicates that students place less importance on the instructional dimension than do teachers, whereas a negative value indicates that students place more importance on the dimension than do teachers. The **rank difference** was obtained by subtracting students' rank from teachers' rank for each characteristic. A positive value indicates that students place more importance on the instructional dimension than do teachers, whereas a negative value indicates that students place less importance on the dimension than do teachers.

These protective measures, coupled with the added advantage of using central limit theorem to meet the assumption of normality of distribution, should provide sufficient confidence in the results of the test. In his discussion of the assumptions of *t*-tests, Field (2009) stresses the importance of sample size in meeting the assumption of normality in *t*-tests. As he puts it:

> …we need to remember that it's the shape of the sampling distribution that matters, not the sample data. One option then is to use a big sample and rely on the central limit theorem which says that the sampling distribution should be normal when samples are big.
>
> (p. 345)

6.1.2.4 Mismatches:

As shown in Table 6.2, the independent sample *t*-test demonstrated a number of significant differences between teachers and students in their perceptions of the differential importance of the various characteristics of good teaching. It can be seen that students rated only five characteristics significantly more important than did the teachers, namely: being friendly (t= -7.18, p<.001), pace of the lecture (t= -9.33, p<.001), building on previous knowledge (t= -8.17, p<.001), intonation & stress (t= -5.80, p<.001), and being a native speaker of the target language (t= -9.16, p<.001).

As far as the pace of lecture for note taking is concerned, the findings in this study confirm the finding of Fisher *et al.* (1998). In their investigation of students' and lecturers' ratings of 21 criteria of good college teaching, they found that pace of lecture was the only criterion students rated significantly more important than did the lecturers. The public speaking skills of the teacher also featured in their study as an important characteristic which received a higher mean rank from the students. This corresponds to the higher ranks received from the students for teacher's intonation

and stress in this study. Without reaching the statistical significance required, it also corresponds to students' higher valuation for teacher's expressiveness found in the present investigation. In Fisher *et al.* (1998), however, "building on students' previous knowledge" was considered significantly more important by the teachers, rather than by the students as is the case in the present study. Being friendly also ranked third in importance by students in a list of 11 "personality" characteristics of excellent teachers in Raymond (2008), while ranked only seventh by the teachers. Saafin (2005) even goes further and concludes that this quality is of high value for Arab students especially in EFL classes where the social context plays an important role in learning. Saafin adds that "The Arab culture values friendliness and considers it as one of the important characteristics of a "good" person" (Saafin, 2005: 88).

On the other hand, teachers rated 16 characteristics significantly more important than did the students. For them, various aspects of group interaction, namely: encouraging students' participation (t= 11.58, p< 001), encouraging questions (t= 6.80, p<.001), classroom management (t= 10.24, p<.001), using student-centred approaches (t= 9.41, p<.001), and inviting ideas (t= 5.64, p<.001) were significantly more important than for the students. The teachers also placed more importance on several aspects of the presentation and facilitation skills of college teaching. In particular, they considered teacher's flexibility and diversity in teaching styles (t= 5.37, p<.001), and presenting different points of view other than lecturer's own (t= 4.48, p<.001) as significantly more important than did the students.

Various teaching dimensions related to the teacher's enthusiasm and rapport with the students were also rated significantly higher by the lecturers. The teacher's

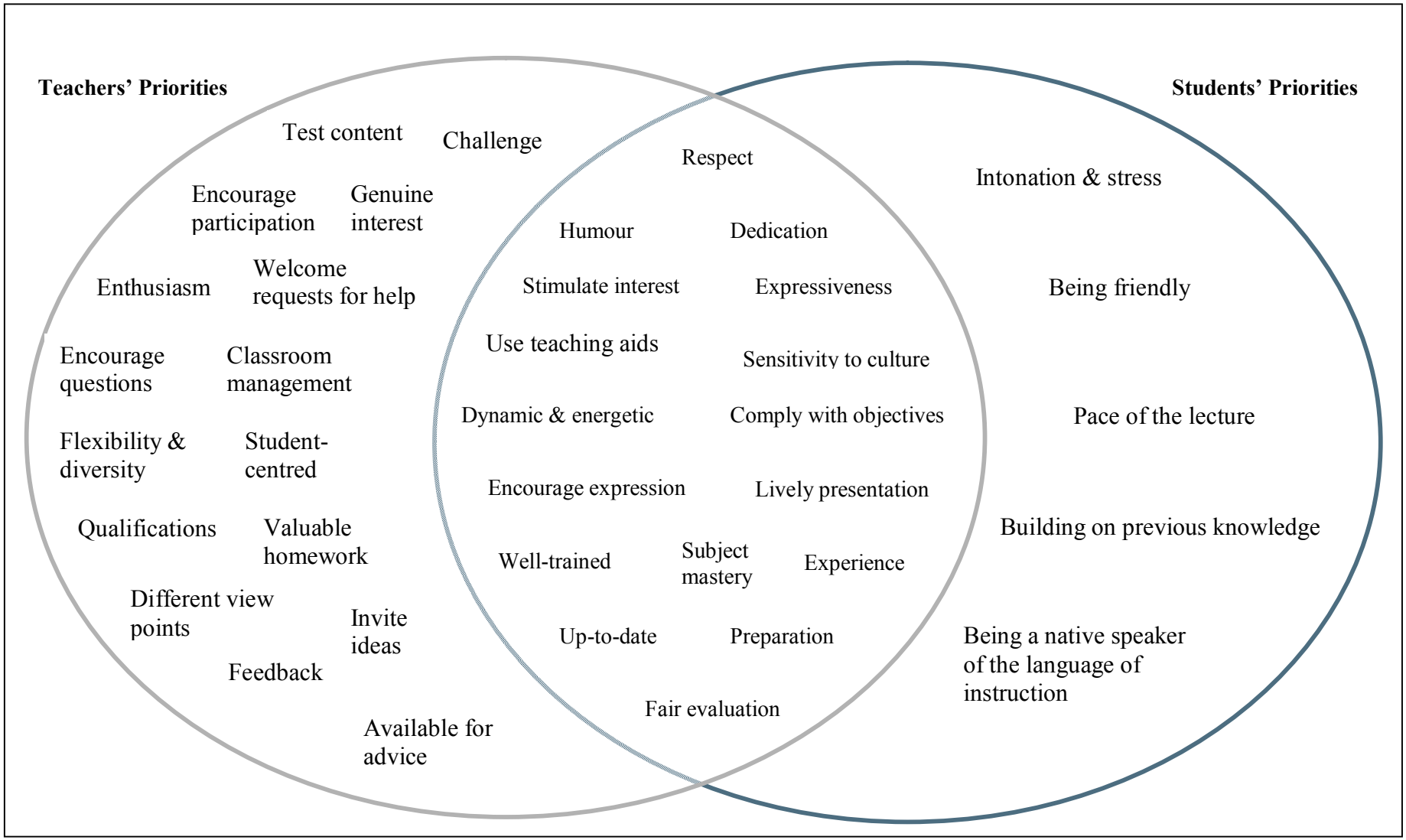
enthusiasm for the subject (t= 4.16, p<.001), making the course intellectually challenging (t= 6.81, p<.001), welcoming requests for help from students (t= 9.62, p< .001), having genuine interest in individual students (t= 4.21, p<.001), and being available to students for advice and support (t= 3.80, p<.001) were all rated as significantly more important by the teachers compared with the students.

Other aspects of teaching where teachers placed higher values are related to the assessment of course work, assignments and homework, and the teacher's academic qualifications. Teachers placed significantly more importance on giving valuable feedback (t=4.51, p<.001), assigning valuable homework (t= 5.92, p< .001), and giving assignments/graded materials which test class content as emphasised by the lecturer (t= 6.70, p<.001). A significant observed difference between teachers' and students' rating of importance was also found in the value the two groups attached to the academic qualifications of the teacher. For teachers, having relevant academic qualifications in teaching English as a second/foreign language (t= 3.52, p<.001) is significantly more important than for the students.

6.1.2.5 Matches:

Despite the 21 statistically significant mismatches reported above, teachers' and students' rank-ordering of the importance of 17 characteristics of effective college teaching seem to be closely matched. Figure 6.2 overleaf summarises all the matches

Figure 6.2: Teachers' and Students' Priorities Compared



and mismatches between the perceptions of the two sample groups with regard to the importance they place on each of the 38 characteristics of effective college teaching. It is clear from Table 6.2 and Figure 6.2 that both teachers and students place a similar level of value on several aspects related to the mastery of the subject matter, teaching experience, and teacher training. There were no significant differences in scores between the teachers and the students for the criteria related to mastery of the subject matter (t= -.35, p= .73), teaching experience (t= .98, p= .33), sufficiency of teacher training (t= -1.89, p= .06), and keeping abreast of the latest developments in the field (t= .19, p= .85).

Although teachers and students differed in their valuation of some aspects of the presentation and facilitation skills as mentioned earlier, they were similar in the importance they attached to other aspects of this dimension of teaching. Both groups regarded preparation of good course materials and carefully explaining them to students as an extremely important characteristic of good college teaching (t= .36, p=.72). Also, no significant differences were found between the two samples' perceptions of the importance of lively presentation styles (t= -1.58, p= .12), and the use of humour in enhancing presentations (t= -.85, p= .39). At a lower level of strength, there were other matches in the perceptions of the two groups with regard to the presentation and facilitation aspect of teaching. Both teachers and students seemed to agree on the weight they assigned to: teacher's expressiveness (t= -2.82, p=.005), teacher's ability to stimulate students' interest in the subject (t= 2.81, p=.005), making good use of teaching aids (t= 2.91, p<.005), and being dynamic and energetic in conducting the class (t= 2.10, p<.05).

No significant differences were also observed between teachers and students in their rank-ordering of the importance of six other characteristics of effective teaching. There were no significant differences in the group scores for dedication to teaching (t= -1.18, p= .24), respect for students (t= -.76, p= .45), fair evaluation (t=-.23, p= .82), encouraging students to express their own ideas (t= -.00, p= .99), showing sensitivity to the culture of the organisation (t= 2.41, p<.05), and complying with the course objectives (t=2.89, p<.005).

6.2 Research question 2: To what extent does students' gender have an effect on their perceptions of the importance of various characteristics of effective college teaching?

A Chi-square test for association between the mediating factor of students' gender and the level of importance they assigned to the 38 characteristics of effective college teaching was carried out (Appendix 10). Again, to guard against Type I error, the alpha level was adjusted to p< .001 instead of the traditional .05 using the Bonferroni adjustment technique mentioned earlier. Therefore, only significant associations at p<.001 are reported in the tables that follow. In order to ensure the robustness of the test, the first two response categories, "not at all important" and "slightly important" were carefully checked for low count cells. This is because Chi-square test requires an expected frequency greater than 5 in each cell to maintain statistical power and robustness (Pallant, 2007; Field, 2009).

Some studies (e.g. Donaldson & Flannery, 1993; Raymond, 2008; Witcher, Onwuegbuzie, and Minor, 2001) reported some differences between male and female students in their rating of various aspects of effective teaching. In the light of these findings, a Chi-square test was carried out to establish whether there is any degree of

association between student gender and his or her perceived ranking of the importance of certain criteria of good teaching in this study.

Unlike Saafin's (2005) conclusion that there were no significant differences between male students and female students in their perceptions of effective EFL teaching, nine characteristics demonstrated a significant deference ($p<.001$) of ranking based on gender in this study. These are: demonstrating a high level of expressiveness and giving clear explanations, giving lectures/tutorials in a style and pace that facilitate note-taking, preparing good course materials and carefully explaining them to students, being able to stimulate the interest of the students in the subject, being dynamic and energetic in conducting the class, using appropriate and fair methods of evaluating student work, demonstrating full compliance with the announced objectives of the course, having full command of the subject matter, and being a native speaker of the target language. Table 6.3 sums up the findings of the Chi-square test for all the 9 dimensions listed above.

Table 6.3: Chi-Square Test Results for Association between "Gender" And Students' Rankings of the Importance of Various Characteristics of Effective Teaching

(Significance level (2-sided) p<.001)

	Rating of Importance					
Gender	**Not at all important**	**Slightly important**	**Moderately important**	**Very important**	**Extremely important**	**χ^2 (Sig.)**
Demonstrating a high level of expressiveness and giving clear explanations						
Male	8	26	57	198	326	**18.491**
	1.3%	4.2%	9.3%	32.2%	53.0%	**(p=.001)**
Female	0	5	18	107	218	
	.0%	1.4%	5.2%	30.7%	62.6%	
Giving lectures/tutorials in a style and pace that facilitate note-taking						
Male	10	32	80	208	282	**38.204**
	1.6%	5.2%	13.1%	34.0%	46.1%	**(p=.000)**

Female	0	4	20	110	216	
	.0%	1.1%	5.7%	31.4%	61.7%	
Preparing good course materials and carefully explaining them to students						
Male	4	16	70	241	284	**33.125 (*p*=.000)**
	.7%	2.6%	11.4%	39.2%	46.2%	
Female	0	1	21	106	221	
	.0%	.3%	6.0%	30.4%	63.3%	
Being able to stimulate the interest of the students in the subject						
Male	6	19	68	171	344	**17.681 (*p*=.001)**
	1.0%	3.1%	11.2%	28.1%	56.6%	
Female	0	3	22	92	231	
	.0%	.9%	6.3%	26.4%	66.4%	
Being dynamic and energetic in conducting the class						
Male	9	18	78	227	279	**31.014 (*p*=.000)**
	1.5%	2.9%	12.8%	37.2%	45.7%	
Female	0	4	17	122	208	
	.0%	1.1%	4.8%	34.8%	59.3%	
Using appropriate and fair methods of evaluating student work						
Male	7	34	83	212	280	**20.353 (*p*=.000)**
	1.1%	5.5%	13.5%	34.4%	45.5%	
Female	2	7	29	108	203	
	.6%	2.0%	8.3%	30.9%	58.2%	
Demonstrating full compliance with the announced objectives of the course						
Male	28	67	150	228	141	**23.344 (*p*=.000)**
	4.6%	10.9%	24.4%	37.1%	23.0%	
Female	4	18	72	159	96	
	1.1%	5.2%	20.6%	45.6%	27.5%	
Having full command of the subject matter						
Male	4	15	63	181	353	**28.694 (*p*=.000)**
	.6%	2.4%	10.2%	29.4%	57.3%	
Female	1	3	16	72	259	
	.3%	.9%	4.6%	20.5%	73.8%	
Being a native speaker of the target language						
Male	91	54	132	153	186	**20.946 (*p*=.000)**
	14.8%	8.8%	21.4%	24.8%	30.2%	
Female	89	37	74	69	81	
	25.4%	10.6%	21.1%	19.7%	23.1%	

It can be seen from Table 6.3 that female students rated 8 out of the 9 dimensions of effective college teaching analysed above higher than their male counterparts, with

the only exception being the nativeness of the teacher to the target language. Starting with the ranking in importance of the ability of the teacher to demonstrate a high level of expressiveness and give clear explanation, 93.3% of the female students participating in this study ranked this dimension of teaching as either extremely important or very important. This compares to 85.2% of the male student population in the study rating this criteria as extremely important or very important.

Another criterion of good teaching which had significant association with student participant's gender was pace of lecture. A vast majority of female students, 93.1%, considered the lecturer's ability to give classes in a suitable pace that facilitates note-taking as either an extremely important or very important quality. This compares to a lower 80.1 % of the male student sample who attached the same level of importance to this characteristic.

Preparing good course materials and carefully explaining them to students also produced a difference of ranking between male and female students. 93.7% of the female students gave the rating extremely important or very important to this quality of effective college teachers. On the other hand, only 85.7% of male students thought of this quality as either extremely important or very important.

Another criterion of effective teaching ranked higher by female students is teacher's ability to stimulate the interest of the students in the subject. This aspect of teaching was ranked very important or extremely important by 92.8% of the female students, compared to 84.7% of the male students.

A fifth area of strong association between participant's gender and importance ranking is lecturer's dynamism and energy in conducting the class. Having a lecturer who is dynamic and energetic in the classroom is far more important for the female students than for the male students. A vast majority of 94.1% of female participants considered it either extremely important or very important for a college lecturer to exhibit this quality. This is compared to a lower 82.9% of male students who gave the same ratings to this dimension of teaching.

The findings in the present investigation that female students tend to attach greater importance to the last two characteristics of effective teaching -*being able to stimulate the interest of the students in the subject, and being dynamic and energetic in conducting the class*- contradicts Raymond's (2008) finding about a related criterion. In her study of Arab students in the Gulf, female students rated the importance of making classes interesting as less important than did the male students.

Using appropriate and fair methods of evaluating students' work was also seen to be extremely important or very important to effective teaching by the vast majority of female students (89.1%). While male students also highly regarded this characteristic of good teachers, their ranking was somewhat lower by around 10%.

Relative to the 6 characteristics discussed above, both sample groups seem to display less enthusiasm for their lecturer's compliance with the course objectives. Most of the students in both groups rated this aspect of teaching as moderately important or very important instead of extremely important. Nevertheless, gender and ranking associations were also observable in this dimension. Here, 66.2% of the female

students' considered this criterion as moderately important or very important, compared to a slightly lower 61.5% of the male students who assigned the same weight to this aspect.

Another characteristic of excellent college teachers where female students differed significantly from their male counterparts in their valuation of its importance was lecturer's command of the subject matter. 73.8% of the female students, compared to only 57.3% of the male students, thought that it was extremely important for a college teacher to have good mastery of his or her subject area.

Finally, male and female students also differed in their perceptions of the importance of their lecturer being a native speaker of the target language. Male students rated this aspect more importantly than female students did. More than half of the male students (55%) considered this quality as either very important or extremely important, while only 42.8% of the female students considered this feature to be very important or extremely important to TESOL teachers. To the contrary, more than one third of the female respondents (36%) thought that it was not important or it was only slightly important for the teacher to be a native speaker of the target language. Although this criterion is mostly applicable in TESOL classes, it also has important implications in courses where the medium of instruction is a second language.

6.3 Research question 3: To what extent do mediating factors such as teachers' ethnic background and mother tongue have and effect on their perceptions of the importance of various characteristics of effective college teaching?

A number of studies have investigated the effect of lecturer's ethnicity and/or mother tongue on students' perceptions of their lecturer's teaching effectiveness (e.g. Finegan & Siegfried, 2000; Ogier, 2003). Also, numerous studies have been carried out on the native versus non-native debate in the ELT industry (e.g. Gill and Rebrova, 2001; Medgyes, 1992; Nayar, 1994; Phillipson, 1996; Ustunluoglu, 2007). Few studies, however, have focused on the preferences and perceptions of teachers and students themselves of the native or non-native teacher of English (Moussu, 2000). This research question attempts to bridge this gap in the research literature and investigates how mediating factors, such as the lecturer's ethnicity and mother tongue, affect how college teachers value certain traits of teaching effectiveness.

A Kruskal-Wallis test for association between the mediating factors of lecturers' ethnic background and mother tongue and the level of importance they assigned to the 38 characteristics of effective college teaching included in this study was carried out. Following the Bonferroni adjustment technique mentioned earlier, the alpha level was adjusted to $p < .001$ (.05/38= .001). Because of the big number of tests carried out, only dimensions of effective teaching with significant associations with the grouping variables are reported in the tables that follow. The full Kruskal-Wallis test results for the 38 dimensions against the mediating factors of lecturer ethnic background and mother tongue, along with the full results of the post-hoc tests, are presented in appendices. References to the numbers of these appendices are made in

the sections below. The effect size of the differences detected by the post-hoc Man-Whitney U tests are assessed using Cohen (1988) criteria of .1=small effect, .3=medium effect, and .5=large effect.

6.3.1 Teacher's Ethnicity as a Mediating Factor

After examining the degree of association between lecturer's ethnicity and lecturers' differential ranking of the importance of the 38 characteristics of effective college teaching using Kruskal-Wallis test (Appendix 11), a significant association was found between lecturers' ethnic backgrounds and lecturers' ranking of the importance of three dimensions of effective teaching, namely: *showing dedication to teaching, keeping abreast of the latest developments in the field, and being a native speaker of the target language.*

6.3.1.1. Teacher's Ethnicity and the Perceived Importance of *Dedication to Teaching* to Effective Teachers

As can be seen in Table 6.4, the Kruskal-Wallis test revealed a statistically significant difference in the perceived importance of *showing dedication to teachin*g between the different ethnic groups in the teacher population, χ^2 (7, n= 245) = 25.691, p=.001.

Table 6.4: Association between Teacher's Ethnic Background and the Perceived Importance of *Showing Dedication to Teaching*

Teacher's Ethnicity	N	Mdn	Mean Rank	χ^2	df	Sig.
Omani Arab	21	5	122.19	25.691	7	.001
Non-Omani Arab	30	5	107.90			
South Asian/Southwest Asian	96	5	140.74			
European	19	4	84.26			
African	15	5	143.00			
North American	41	4	104.90			
Southeast Asian	17	5	129.74			
Other	6	4.5	94.67			
Total	245					

The teachers from African ethnic backgrounds seem to value *dedication to teaching* the most with a mean rank of 143. With a slightly lower mean rank of 140.74, South Asian and Southwest Asian teachers come second. European teachers, on the other hand, were the least enthusiastic for this aspect of teaching with a mean rank of only 84.26.

In light of the statistically significant differences in the perceptions held by the different ethnic groups in the teacher population with regard to the perceived importance of *showing dedication to teaching* to effective teachers revealed by this test, multiple post-hoc Mann-Whitney U tests were carried to examine this relationship further. These follow-up tests were conducted between pairs of ethnic groups to know which of the groups are statistically significantly different from one another. Each group was compared with one another, resulting in 28 comparisons. As a result the, a Bonferroni correction to the alpha value was applied (.05/28= .002). The effect size (*r*) of the differences detected by the post-hoc Man-Whitney U tests are assessed using Cohen (1988) criteria of .1=small effect, .3=medium effect, and .5=large effect. The approximate value of *r* was calculated using the following formula recommended by Pallant (2007: 223):

r = z/square root of N, where N= total number of cases in each test

Of the 28 tests, only two revealed statistically significant differences between two pairs of ethnic groups. These two tests are presented in Table 6.5. The rest of the post-hoc tests are presented in Appendix 12.

Table 6.5: Post-Hoc Tests for the Association between Teacher's Ethnic Background and the Perceived Importance of *Showing Dedication to Teaching*

Teacher's Ethnicity	N	Mdn	Mean Rank	*U*	*Z*	Sig.	*r*
South Asian/ Southwest Asian	96	5	62.27	502.500	-3.937	0.000	.37
European	19	4	36.45				
South Asian/ Southwest Asian	96	5	75.13	1380.000	-3.461	0.001	.30
North American	41	4	54.66				

As can be seen from Table 6.5, South Asian and Southwest Asian teachers appear to attach more importance to *showing dedication to teaching* compared to their European and North American colleagues. South Asian and Southwest Asian teachers (*n*= 96, *Mdn*=5, *Mean Rank*= 62.27) gave a much higher ranking to this characteristic of effective teaching compared to European teachers (*n*=19, *Mdn*=4, *Mean Rank*=36.45), *U*= 502.500, *z*=-3.937, *p*=.000, *r*=.37. When compared to the North American teachers (*n*=41, *Mdn*=4, *Mean Rank*=54.66), South Asian and Southwest Asian teachers (*n*=96, *Mdn*= 5, *Mean Rank*= 75.13) also appeared to attach greater importance to dedication to the teaching profession as a characteristic of effective teachers, *U*=1380.000, *z*=-3.461, *p*=.001, *r*=.30.

6.3.1.2 Teacher's Ethnicity and the Perceived Importance of *Keeping Abreast of the Latest Developments in the Field* to Effective Teaching

A Kruskal-Wallis test was carried out to investigate the effect of teachers' ethnic backgrounds on their perception of the importance of *keeping abreast of the latest developments in the field* (Table 6.6). The test revealed a statistically significant difference between the rankings given by the different ethnic groups to the importance of this characteristic of effective teaching, χ^2 (7, *n*= 243) = 27.088, *p* = .000.

Table 6.6: Association between Teacher's Ethnic Background and the Perceived Importance of *Keeping Abreast of the Latest Developments in the Field*

Teacher's Ethnicity	N	Mdn	Mean Rank	χ^2	df	Sig.
Omani Arab	21	5	145.38	27.088	7	.000
Non-Omani Arab	30	4	128.53			
South Asian/Southwest Asian	95	4	128.92			
European	19	4	97.95			
African	15	4	144.67			
North American	40	4	86.28			
Southeast Asian	17	4	151.18			
Other	6	3.5	72.92			
Total	243					

As can be seen in Table 6.6, teachers from Southeast Asia and from Oman assigned the highest ranks to *keeping abreast of the latest developments in the field*, with a Mean Rank of 151.18 and 145.38 respectively. On the other hand, European and North American teachers gave much lower ranks to this aspect of effective teaching, a Mean Rank of 97.95 and 86.28 respectively.

This statistically significant difference in the perceptions of the different teacher ethnic groups of the importance of *keeping abreast of the latest developments in the field* necessitated a follow-up test to determine which groups of teachers differed from one another the most. Again, post-hoc Mann-Whitney U tests were carried out between pairs of ethnic groups to know which of the groups are statistically significantly different from one another, resulting in 28 post-hoc tests. As a result the, the Bonferroni correction to the alpha value was applied (.05/28= .002). Two post-hoc tests revealed statistically significant differences between two pairs of ethnic groups. The results of these two tests are presented in Table 6.7. The rest of the post-hoc tests for this dimension are presented in Appendix 13.

Table 6.7: Post-Hoc Tests for the Effect of Teacher's Ethnic Background on the Perceived Importance of *Keeping Abreast of the Latest Developments in the Field*

Teacher's Ethnicity	N	Mdn	Mean Rank	*U*	*Z*	Sig.	*r*
South Asian/ Southwest Asian	95	4	75.28	1208.000	-3.619	0.000	.31
North American	40	4	50.70				
North American	40	4	24.48	159.000	-3.376	0.001	.45
Southeast Asian	17	4	39.65				

As shown in Table 6.7, South Asian and Southwest Asian teachers (*n*= 95, *Mdn*=4, *Mean Rank*= 75.28) appear to attach greater significance to *keeping abreast of the latest developments in the field* compared to their North American colleagues (*n*=40, *Mdn*=4, *Mean Rank*=50.70), *U*= 1208.000, *z*=-3.619, *p*=.000, *r*=.31. North American teachers also seem to place less weight on this trait of effective teachers (*n*=40, *Mdn*=4, *Mean Rank*=24.48) compared to their Southeast Asian counterparts (*n*=17, *Mdn*= 4, *Mean Rank*= 39.65), *U*=159.000, *z*=-3.376, *p*=.001, *r*=.45.

As far as professional development is concerned, there could be a number of reasons behind Asian teachers' emphasis on the importance of *keeping abreast of the latest developments in the field,* especially when examining this in a TESOL context. The vast majority of Asian TESOL teachers are non-native speakers of English. As stressed by Maum (2002) the credibility of non-native English speaking teachers in TESOL classes is being challenged by what she termed the "inevitable trickle-down effect of the native speaker fallacy" (Maum, 2002: 1). Students sometimes resent being taught by non-native speakers of English, and this puts the non-native speaking teachers under an additional pressure to try to assert themselves in the profession as competent, well-qualified, teachers with up-to-date knowledge of the subject matter. As Ustunluoglu (2007: 71) puts it:

> It might be assumed that a non-native teacher needs to study more than a native teacher does as she/he teaches a language which is not her/his native language and this may lead him/her to study harder, come better prepared and informed about the subject.

It is difficult to tell whether North American TESOL teachers working in the GFP programs assign less importance to *keeping abreast of the latest developments in the field* because they are less enthusiastic for professional development and less committed to keeping up-to-date in their field or because they feel more confident of their linguistic abilities and skills in a TESOL context where many of the teachers are non-native speakers of English. Either way, if this perception is translated into a real life attitude that does not recognise the importance of keeping up-to-date with the latest developments in the field, this attitude may ultimately lead to conflict or tension with programs managers from a different ethnic background or with a different view of the world.

6.3.1.3 Teacher's Ethnicity and the Perceived Importance of *Being a Native Speaker of the Target Language* to Effective Teaching

To investigate the effect of teachers' ethnic backgrounds on their ranking of the importance of *being a native speaker of the target language* to effective teaching, a Kruskal-Wallis test was carried out (Table 6.8). The test revealed a statistically significant difference between the weights assigned to this characteristic by the teachers from different ethnic groups, χ^2 (7, n= 243) = 36.801, p = .000.

Table 6.8: The Effect of Teacher's Ethnic Background on the Perceived Importance of *Being a Native Speaker of the Target Language* to Effective Teaching

Teacher's Ethnicity	N	Mdn	Mean Rank	χ^2	df	Sig.
Omani Arab	21	3	127.02	36.801	7	.000
Non-Omani Arab	30	2.5	123.38			
South Asian/Southwest Asian	95	2	94.95			

European	19	3	153.16			
African	15	3	132.07			
North American	40	3	166.31			
Southeast Asian	17	3	115.18			
Other	6	2	125.83			
Total	243					

As evident from Table 6.8, North American teachers (Mean Rank 166.31) and European teachers (Mean Rank 153.16) recorded the highest valuation for *being a native speaker of the target language* to effective teaching. South Asian and Southwest Asian teachers, however, seem to place much less importance on this trait (Mean Rank 94.95).

The statistically significant difference in the Kruskal-Wallis test reported above called for post-hoc tests to identify the pairs of ethnic groups with the most significant differences in the perceived importance of *being a native speaker of the target language* to effective teaching. Mann-Whitney U tests were carried out between all the possible 28 combinations of ethnic groups. To protect the alpha level, the Bonferroni correction was applied (.05/28= .002). Three combinations revealed statistically significant differences. These are presented in Table 6.9 below. The rest of the post-hoc tests are presented in Appendix 14.

Table 6.9: Post-Hoc Tests for the Effect of Teacher's Ethnic Background on the Perceived Importance of *Being a Native Speaker of the Target Language* to Effective Teaching

Teacher's Ethnicity	N	Mdn	Mean Rank	*U*	*Z*	Sig.	*r*
Non-Omani Arab	30	2.5	27.27	353.000	-3.032	0.002	.36
North American	40	3	41.68				
South Asian/Southwest Asian	95	2	52.78	454.000	-3.591	0.000	.34
European	19	3	81.11				
South Asian/Southwest Asian	95	2	56.76	832.500	-5.371	0.000	.46
North American	40	3	94.69				

As shown in Table 6.9, lecturers from North America and Europe teaching in the GFP TESOL programs surveyed in this study seem to attach more importance to the nativeness of the lecturer to the target language (English) compared to their Arab and Asian colleagues. North American teachers (*n*=40, *Mdn*=3, *Mean Rank*=41.68) ranked this characteristic of effective teachers higher than the non-Omani Arab teachers (*n*=30, *Mdn*=2.5, *Mean Rank*=27.27), *U*= 353.000, *z*=-3.032, *p*=.002, *r*=.36. An even greater difference was found between North American teachers (*n*=40, *Mdn*=3, *Mean Rank*=94.69) and South Asian and Southwest Asian teachers (*n*=95, *Mdn*=2, *Mean Rank*=56.76), *U*=832.500, *z*=-5.371, *p*=.000, *r*=.46. European teachers (*n*=19, *Mdn*=3, *Mean Rank*=81.11) were also found to assign more weight to *being a native speaker of the target language* compared to the teachers from South Asia or Southwest Asia (*n*=95, *Mdn*=2, *Mean Rank*=52.78).

As far as this aspect of lecturing is concerned, North American and European lecturers' valuation of this characteristic of teachers is closer to students' priorities discussed in section 6.1.2. This raises questions on how lecturers from western countries working as TESOL teachers in Oman and the Gulf may respond to teacher appraisal schemes managed by Arab or south-western Asian colleagues who do not highly regard nativeness to English as an important aspect of teaching effectiveness. It also raises questions about the effects on work environment and intercultural relationships between lecturers from different ethnic backgrounds this difference in perception may cause, especially that the majority of English teachers in the world are not native speakers of English (Matsuda & Matsuda, 2001).

6.3.2 Teacher's Mother Tongue (L1) as a Mediating Factor

Examining teachers' differential ranking of the importance of the 38 characteristics of effective college teaching against the teachers' mother tongues, a Kruskal-Wallis test revealed two significant differences (Appendix 15). Significant differences were observed between the rankings given by the teachers from different mother tongue groups in two areas of teaching effectiveness, namely: *having relevant academic qualifications in teaching English as a second/foreign language*, and *being a native speaker of the target language.*

6.3.2.1 Teachers' Mother Tongues (L1) and the Perceived Importance of *Having Relevant Academic Qualifications in Teaching English as a Second/Foreign language*

To test for differences between the rankings given by the different L1 groups of teachers to the importance of *having relevant academic qualifications in teaching English as a second/foreign language* to effective teaching in TESL/TEFL classes, a Kruskal-Wallis test was carried out (Table 6.10). The test revealed a statistically significant difference between the rankings assigned to this characteristic by the teachers from different L1 groups, χ^2 (10, n= 242) = 37.766, p = .000.

Table 6.10: The Effect of Teacher's Mother Tongue on the Perceived Importance of *Having Relevant Academic Qualifications in Teaching English as a Second/Foreign Language* to Effective Teaching

Teacher's Mother Tongue	N	Mdn	Mean Rank	χ^2	df	Sig.
Arabic	49	5	145.28	37.766	10	0.000
English	75	4	94.73			
Urdu	14	4	92.89			
Hindi	18	4	104.03			
Malayalam	26	5	164.54			
Tamil	18	4	124.47			
Tagalog	17	4	118.85			
Other South Asian/ Southwest Asian language	14	5	137.39			
African language	5	4	120.00			

European language	4	5	141.75			
Other	2	5	88.50			
Total	242					

As evident from the data in Table 6.10, L1 speakers of Malayalam (Mean Rank 164.54), Arabic (Mean Rank 145.28), and European languages (Mean Rank 141.75) highly ranked the importance of academic qualifications in teaching English as a second/foreign language as a trait of effective college teachers in TESOL programs. On the other hand, teachers who spoke English or Urdu as their L1 gave a much lower ranking to this characteristic, Mean Ranks 94.73 and 92.89 respectively.

A follow-up test was needed to establish the degree of similarity/difference between the different pairs of L1 speakers. Mann-Whitney U tests were carried out between all the possible 55 combinations of mother tongue groups. To protect the alpha level, the Bonferroni correction was applied (.05/55= .001). Three combinations revealed statistically significant differences. These are presented in Table 6.11 below. The rest of the results for this post-hoc test are presented in Appendix 16.

Table 6.11: Post-Hoc Tests for the Effect of Teacher's Mother Tongue on The Perceived Importance of *Having Relevant Academic Qualifications in Teaching English as a Second/Foreign Language* to Effective Teaching

Teacher's Mother Tongue	N	Mdn	Mean Rank	*U*	*Z*	Sig.	*r*
Arabic	49	5	77.77				
English	75	4	52.53	1089.500	-4.052	0.000	.36
English	75	4	44.03				
Malayalam	26	5	71.10	452.500	-4.293	0.000	.43
Urdu	14	4	12.61				
Malayalam	26	5	24.75	71.500	-3.534	0.000	.56

From Table 6.11, it can be seen that teachers who spoke Arabic or Malayalam as their mother tongue gave significantly higher rankings to *having relevant academic*

qualifications in teaching English as a second/foreign language compared to their colleagues who spoke English or Urdu as their first language. L1 speakers of Arabic (*n*=49, *Mdn*=5, *Mean Rank*=77.77) differed significantly from the teachers who spoke English as their first language (*n*=75, *Mdn*=4, *Mean Rank*=52.53) in their valuation of the importance of academic qualifications to effective teaching in TESOL, *U*= 1089.500, *z*=-4.052, *p*=.000, *r*=.36. English L1 speakers (*n*=75, *Mdn*=4, *Mean Rank*=44.03) also gave a significantly lower ranking of academic qualifications compared to the L1 speakers of Malayalam (*n*=26, *Mdn*=5, *Mean Rank*=71.10), *U*= 452.500, *z*=-4.293, *p*=.000, *r*=.43. Teachers whose mother tongue was Malayalam (*n*=26, *Mdn*=5, *Mean Rank*=24.75) also placed more importance on academic qualifications than the L1 speakers of Urdu (*n*=14, *Mdn*=4, *Mean Rank*=12.61), *U*= 71.500, *z*=-3.534, *p*=.000, *r*=.56.

From the post-hoc tests shown above in Table 6.11 and the other post-hoc tests nearing the significance level set for these tests shown in Appendix 16, it can be said that non-native speakers of English in the teacher population surveyed in this study tend to place more importance on *having relevant academic qualifications in teaching English as a second/foreign language* compared to the native speakers of English. This could be attributed to a number of reasons. One explanation could be that Arab countries and countries of the Indian sub-continent are developing countries in which academic qualifications, as opposed to experience, apprenticeship, or in-service training, are still regarded as the single most important asset for job seekers, especially in the public sector. In addition, TESOL lecturers whose first language is English are in a better position to secure better employment opportunities in ELT, even in the absence of proper academic qualifications (Amin, 2000; Braine,

1999; Canagarajah, 1999; Maum, 2002; Rampton, 1996). However, "People do not become qualified to teach English merely because it is their mother tongue, and much of the knowledge that native speakers bring intrinsically to the ESL classroom can be learned by [non-native English speaking teachers] through teacher training" (Maum, 2002: 1).

6.3.2.2 Teachers' Mother Tongues (L1) and the Perceived Importance of *Being a Native Speaker of the Target Language*

A Kruskal-Wallis test was carried out to establish whether there were statistically significant L1-based differences between the teachers in their perceptions of the importance of *being a native speaker of the target language* to effective teaching (Table 6.12). The test revealed a statistically significant difference between the rankings given by the different L1 groups to the importance of this characteristic of effective teaching, χ^2 (10, n= 241) = 55.583, p = .000. In a way, this strong association resembles the strong relationship found between lecturer's ethnicity and lecturer's rating of this same trait in section 6.3.1.3.

Table 6.12: The Effect of Teacher's Mother Tongue on The Perceived Importance of *Being A Native Speaker of the Target Language* to Effective Teaching

Teacher's Mother tongue	N	Mdn	Mean Rank	χ^2	*df*	Sig.
Arabic	49	3	127.04	55.583	10	0.000
English	74	3	160.75			
Urdu	15	3	116.20			
Hindi	17	2	109.26			
Malayalam	26	1	65.48			
Tamil	18	1	83.39			
Tagalog	17	3	114.00			
Other South Asian or Southwest Asian language	14	1.5	87.86			
African language	5	1	86.20			
European language	4	2	97.50			
Other	2	2.5	123.75			
Total	241					

As displayed in Table 6.12, by far, native speakers of English (Mdn=3, Mean Rank= 160.75) assigned the highest ranking to the importance of *being a native speaker of the target language* among the different L1 groups of teachers. On the other hand, the teachers who spoke Malayalam as their first language (Mdn= 1, Mean Rank= 65.48) gave much lower rankings to this characteristic of TESOL teachers.

This statistically significant difference in the perceptions of the different L1 teacher groups of the importance of *being a native speaker of the target language* to teaching effectiveness required a follow-up test to determine which groups of teachers differed significantly from one another. Post-hoc Mann-Whitney U tests were carried out between all the possible 55 pairs of L1 combinations. A Bonferroni correction to the alpha value was applied (.05/55= .001). Four post-hoc tests revealed statistically significant differences between four pairs of L1 groups. The results of these four tests are presented in Table 6.13. The rest of the post-hoc tests for this dimension are presented in Appendix 17.

Table 6.13: Post-Hoc Tests for the Effect of Teacher's Mother Tongue on the Perceived Importance of *Being a Native Speaker of the Target Language* to Effective Teaching

Teacher's Mother tongue	N	Mdn	Mean Rank	*U*	*Z*	Sig.	*r*
Arabic	49	3	44.94				
Malayalam	26	1	24.92	297.000	-3.973	0.000	.46
English	74	3	60.28				
Malayalam	26	1	22.67	238.500	-5.839	0.000	.58
English	74	3	51.93				
Tamil	18	1	24.19	264.500	-4.061	0.000	.42
English	74	3	48.61				
Other South Asian/ Southwest Asian language	14	1.5	22.75	213.500	-3.564	0.000	.38

As can be seen in Table 6.13, in three out of four tests, native speakers of English seem to place greater importance to *being a native speaker of the target language* (English) compared to the speakers of three other L1 groups, namely: Malayalam, Tamil, and other South Asian and Southwest Asian Languages. Native speakers of English (*n*=74, *Mdn*=3, *Mean Rank*= 60.28) were found to be stronger advocates of this criterion compared to their colleagues whose first language was Malayalam (*n*=26, *Mdn*=1, *Mean Rank*= 22.67), *U*= 238.500, *z*=-5.839, *p*=.000, *r*=.58. The teachers who spoke English as their first language (*n*=74, *Mdn*=3, *Mean Rank*= 51.93) were also markedly different in their valuation of this criterion from the natives of Tamil (*n*=18, *Mdn*=1, *Mean Rank*= 24.19), *U*= 264.500, *z*=-4.061, *p*=.000, *r*=.42. Moreover, the rankings of the English L1 speakers (*n*=74, *Mdn*=3, *Mean Rank*= 48.61) compared favourably with the rankings of the L1 speakers of other South Asian and Southwest Asian languages (*n*=14, *Mdn*=1.5, *Mean Rank*= 22.75), *U*= 213.500, *z*=-3.564, *p*=.000, *r*=.38.

The tendency by native English speaking teachers to underestimate the strengths and contributions of non-native English speaking teachers to ELT has been criticised and contested by many in the last twenty years. Nayar (1994: 4) criticises the native-nonnative paradigm and cautions that "Sociolinguists have long pointed out our tendencies to evaluate people through their language, but applied linguists have not yet woken up to our tendency to evaluate the language through the people". Phillipson (1992) goes even further and claims that EFL and ESL are nothing but mere "commodities" created by the native speaker countries for the disempowerment of the other English speakers.

It is the current researcher's position and the position of numerous others (e.g. Bennett, 1994; Gill & Rebrova, 2001; Medgyes, 1992) that a more positive approach is needed in addressing the native/non-native issue. As argued by Medgyes (1992: 349):

> …the ideal [native English speaking teacher] and the ideal [non-English speaking teacher] arrive from different directions but eventually stand quite close to one another…In an ideal school, there should be a good balance of NESTs and non-NESTs, who complement each other in their strengths and weaknesses.

Langage programs need the intrinsic advantages both groups bring to the classroom. The native speaker provides a valuable opportunity for the student to learn the language from its natives and observe and appreciate the cultural and social values embedded in it. The non-native speaker of English TESOL teacher, on the other hand, being a second language learner of the target language himself/herself, can bring a lot of valuable experiences to the classroom that may help students overcome the difficulties of learning a second language.

6.4 Chapter Summary

Using the findings from the research literature and the findings of the qualitative exploratory study, a ranked scale comprising 38 characteristics of effective college teaching was constructed. A sample of 968 students and 248 college teachers from the GFP program in six colleges of technology in Oman were asked to rank the importance of each characteristic on the scale given, 1= Not at all important to 5= Extremely important.

Teachers' and students' rank-ordering of the importance of the 38 characteristics of effective teaching were then correlated. A moderate but statistically significant overall correlation, *rho*= .56, $p<.001$, was found between lecturers' and students'

differential ranking. This overall correlation strength is slightly lower than what was reported in some benchmark studies in the field.

After comparing the group means and ranks, students' and teachers' differential rankings of the 38 characteristics were subjected to an independent sample *t*-test to identify any statistically significant matches/mismatches in the perceived level of importance between the two groups. Students and teachers significantly differed in the rankings of the importance of 21 characteristics of effective teaching, but were closely matched on the other 17 characteristics.

Five characteristics were significantly more important for the students than they were for the teachers. These were: being friendly, pace of the lecture, building on previous knowledge, intonation & stress, and being a native speaker of the target language.

On the other hand, teachers ranked 16 characteristics significantly more important than did the students. These were: encouraging students' participation, encouraging questions, classroom management, using student-centred approaches, inviting ideas, flexibility and diversity in teaching styles, presenting different points of view other than lecturer's own, teacher's enthusiasm for the subject, making the course intellectually challenging, welcoming requests for help from students, having genuine interest in individual students, and being available to students for advice and support, giving valuable feedback, assigning valuable homework, giving assignments/graded materials which test class content as emphasised by the lecturer, and having relevant academic qualifications in teaching English as a second/foreign language.

However, teachers' and students' rank-ordering of the importance of the other 17 characteristics of effective college teaching seemed to be closely matched. There were no significant differences between the teachers and the students in the perceived importance of: mastery of the subject matter, teaching experience, sufficiency of teacher training, and keeping abreast of the latest developments in the field, preparation of good course materials and carefully explaining them to students, using lively presentation styles, using humour to enhance presentations, teacher's expressiveness, teacher's ability to stimulate students' interest in the subject, making good use of teaching aids, being dynamic and energetic in conducting the class, dedication to teaching, respect for students, fair evaluation, encouraging students to express their own ideas, showing sensitivity to the culture of the organisation, and complying with the course objectives.

The next level of analysis was to test for differences in perceptions of effective teaching between groups based on some background variables. For the students, nine traits of good teachers demonstrated a significant deference in ranking based on student's gender. These were: demonstrating a high level of expressiveness and giving clear explanations, giving lectures/tutorials in a style and pace that facilitate note-taking, preparing good course materials and carefully explaining them to students, being able to stimulate the interest of the students in the subject, being dynamic and energetic in conducting the class, using appropriate and fair methods of evaluating student work, demonstrating full compliance with the announced objectives of the course, having full command of the subject matter, and being a native speaker of the target language. Female students rated 8 out of these 9

dimensions of effective college teaching higher than their male counterparts, the only exception being the nativeness of the teacher to the target language

As for the teachers, significant differences were found between the different ethnic groups in the teacher population in their ranking of the importance of three dimensions of effective teaching, namely: *showing dedication to teaching, keeping abreast of the latest developments in the field, and being a native speaker of the target language.* South Asian and Southwest Asian teachers appeared to attach more importance to *showing dedication to teaching* compared to their European and North American colleagues. South Asian, Southwest Asian, and Southeast Asian teachers also attached greater significance to *keeping abreast of the latest developments in the field* compared to their North American colleagues. North American and European teachers in the GFP TESOL programs, however, seem to attach more importance to the nativeness of the lecturer to the target language (English) compared to their Arab and Asian colleagues.

Significant differences were also observed between the perceptions of the different L1 groups in the teacher population, specifically in two dimensions of teaching effectiveness: *having relevant academic qualifications in teaching English as a second/foreign language*, and *being a native speaker of the target language.* Teachers who spoke Arabic or Malayalam as their mother tongue assigned significantly higher rankings to *having relevant academic qualifications in teaching English as a second/foreign language* compared to their colleagues who spoke English or Urdu as their first language. Native speakers of English, however, placed greater importance on *being a native speaker of the target language* (English)

compared to the speakers of Malayalam, Tamil, or other South Asian and Southwest Asian Languages.

CHAPTER SEVEN

TEACHERS' AND STUDENTS' PERCEPTIONS OF STUDENTS' EVALUATIONS OF COLLEGE TEACHING, THEIR HYPOTHESISED BIASING FACTORS, THEIR UTILITY, AND THE ROLE OF THE STUDENT AS AN EVALUATOR OF TEACHING EFFECTIVENESS

7.0 Introduction

In chapter six, the findings about teachers' and students' perceptions of the importance of various characteristics to effective college teaching and the mediating factors that may affect their perceptions were presented and discussed. In this chapter the perceptions and views of the same two sample groups will again be compared and contrasted, but with regard to students' evaluations of teaching (SET) or students' ratings. This chapter takes the discussion to another level to include the findings on the similarities and differences found between GFP students' and teachers' perceptions of: the hypothesised biasing factors in SET, SET utility, and the role of the student as an evaluator of teaching.

As pointed out in Chapter 4, despite the huge body of research literature on student ratings, relatively little research has been done to investigate how students perceive SETs (Kwan, 2000; Sojka, Gupta, and Deeter-Schmelz, 2002; Young *et al.*, 1999). Kwan (2000) asserts that it is particularly important to know and understand students' perspectives on the subject, as without this knowledge, "it is unlikely that we will be able to interpret and use the data [from SETs) in a sensible and valid manner" (p.2). Sojka *et al.* (2002) argue it further and stress the importance of

investigating students' views about SET, especially when teachers' criteria in judging effective teaching may differ from those of the students:

> Because students are the ones who complete the SET and who are the most likely to benefit from SET information via improved instruction in the classroom, it is important that their perceptions on SET also be considered. … Because students' definition of effective teaching may differ from that of faculty members, we need to evaluate the factors influencing SET from the students' points of view, and compare them to faculty perceptions.
>
> (pp. 44-45)

It is now evident from the findings in Chapter six that differences between teachers and students in their perceptions of the qualities of effective teachers do exist. This calls for a further investigation to establish whether teachers and students also differ in their perception of student evaluation of teaching. To this end, data collected from section two of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire will be presented and discussed in an attempt to answer one main research question:

Research question 4: To what extent do teachers' and students' perceptions of students' evaluations of college teaching match or mismatch on:

- **The effect of the hypothesised biasing factors on students' ratings?**
- **The utility of students' evaluations of college teaching?**
- **The role of the student as an evaluator of teaching effectiveness?**

7.1 Research question 4: To what extent do teachers' and students' perceptions of students' evaluations of college teaching match or mismatch?

Section two in the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire explored research participants' perceptions on three dimensions of SETs: the hypothesised biasing factors in SET (11 items), SET utility (4 items), and the role of the student as an evaluator of teaching (4 items). On a

Likert-type scale of four response categories, 1= strongly disagree, 2= disagree, 3= agree, and 4=strongly agree, respondents were asked to indicate the level of their agreement or disagreement to each item. To optimise the reliability of the scale and guard against any data contamination or respondent predisposition that may result from item category labels, clustering of items, or the order in which categories are presented, the 19 items were mixed up and presented in one scale without grouping them into categories or using category labels for subscales.

The Statistical Package for Social Sciences (SPSS), version 15, and Microsoft Excel 2003 were used to analyse the data resulting from this section of the questionnaire. Item mean was used to compare the perceptions of the two sample groups, students (N= 968) and teachers (N= 248). Independent Sample *t*-test was used to identify statistically significant differences in the means between teachers and students. Again, bearing in mind the big number of variables involved in each test, the alpha value was adjusted to a more conservative level using the Bonferroni adjustment technique and was set to $p < .003$ (.05/19= .003), instead of the traditional .05 or .01 levels. As mentioned above, the purpose of the first 11 items on the scale is to probe research participants' views about the factors believed to bias SET. The findings on this dimension are presented in the following section.

7.1.1 Teachers' and Students' Perceptions of the Effect of the Hypothesised Biasing Factors on Students' Ratings

As discussed in Chapter 4, section 4.5, various factors unrelated to teaching effectiveness have been hypothesised to bias students' ratings in the SET research literature. While many of them have been dismissed as pure myths by some researchers in the field (Aleamoni, 1987, Aleamoni, 1999, Feldman, 1997), a number

of them are frequently cited in SET research as potential threats to the validity and utility of students' ratings. Researchers in the field generally group these biasing variables under four categories:

- Course characteristics
- Student characteristics
- Instructor characteristics
- Administrative procedures & rating instrumentation

Table 7.1 lists 11 hypothesised factors that are believed to bias students' ratings and that were included in section two of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire. For every hypothesised factor, a statement was presented to probe the views and perceptions of the research participants about the validity of students' ratings. As mentioned earlier in Chapter 5, these hypothesised factors were derived from the literature review as well as from the findings of the exploratory study. For each statement, an abbreviation is given in Table 7.1 for easy reference and presentation in the tables, figures and discussions that follow.

Table 7.1 Hypothesised Factors Affecting the Validity of Students' Ratings and Their Abbreviations

	Hypothesised factors affecting validity of SET	Abbreviation
1.	A student's prior interest in the subject affects his/her rating of this subject's instructor	Student's prior interest
2.	Students give lower ratings to teachers of courses with high workload	Course workload
3.	Lecturer's personal attributes play a major role in students' ratings	Lecturer's personal attributes
4.	Creative college lecturers may be poorly rated by students if these lecturers are unaware of the learning styles students are used to in pre-college education	Students' preferences of teaching style
5.	I feel that students' rating of/reaction to their lecturer's teaching is significantly influenced by the fact that s/he is a native/non-native speaker of English	Lecturer's mother tongue

6.	Student evaluation of teaching can cause lecturers to deflate course workload and lower standards in order to keep students happy	Workload deflation & lowering of standards
7.	The grade students expect in a course affects how they rate their lecturer in that course	Expected grade
8.	Students' ratings of lecturers for making personnel decisions such as contract renewal or termination is a main cause of grade inflation	Grade inflation
9.	Lecturers are more likely to receive high students' ratings when evaluated by students of the opposite sex	Gender
10.	Lecturers with good communication skills are more likely to receive high students' ratings	Lecturer's Communication skills
11.	Students' judgment of the teaching performance of a lecturer can be affected by the lecturer's ethnic background or nationality	Lecturer's ethnicity

Teachers' and students' perceptions of the validity of students' ratings of college teachers were first compared by calculating the item mean. A high mean value in this category of items would indicate a stronger effect for the hypothesised factor on the SET validity, while a low mean would indicate a weaker effect for the hypothesised factor on the SET validity as perceived by the research participants. The results of the independent sample *t*-test were used to identify significant differences between the means of the two sample groups. Levene's test of equality of variances was used to identify the most suitable t-value and its associated significance level of differences between group means for each dependent variable. These means and the *t* test results for each item are compared for teachers and students in Table 7.2.

Table 7.2: Teachers' and Students' Perceptions of the Validity of Students' Ratings Compared

	Teachers					Students						
	N	**Min**	**Max**	**M**	**SD**	**N**	**Min**	**Max**	**M**	**SD**	***t***	**Sig.**
Student's prior interest	245	1	4	2.93	.752	963	1	4	2.97	.870	-.706	.480
Course workload	241	1	4	2.66	.779	963	1	4	2.29	.898	6.494	**.000**
Lecturer's personal attributes	243	1	4	3.18	.660	963	1	4	3.35	.744	-3.471	**.001**
Students' preferences of teaching style	240	1	4	2.96	.751	939	1	4	2.80	.865	2.826	.005
Lecturer's mother tongue	242	1	4	2.42	.913	959	1	4	2.89	.909	-7.087	**.000**
Workload deflation & lowering of standards	240	1	4	2.66	.877	959	1	4	2.50	.957	2.496	.013
Expected grade	244	1	4	2.88	.817	943	1	4	2.77	.858	1.790	.074
Grade inflation	237	1	4	2.58	.858	960	1	4	2.41	.955	2.699	.007
Gender	241	1	4	2.22	.755	959	1	4	2.37	1.019	-2.666	.008
Lecturer's Communication skills	245	1	4	3.11	.638	950	1	4	3.13	.851	-.474	.635
Lecturer's ethnicity	242	1	4	2.62	.894	964	1	4	2.25	1.100	5.584	**.000**

*** Significant differences at $p < .003$ (2-tailed)**

The mean difference and the corresponding *t* value with its significance level for the individual items indicate a varying degree of similarity/ difference between teachers' and students' perception of the effect of the factors hypothesised to bias students' ratings. Teachers significantly differ from students in their assessment of the impact of four factors on the validity of SET, namely: course workload, lecturer's personal attributes, lecturer's mother tongue, and lecturer's ethnic background.

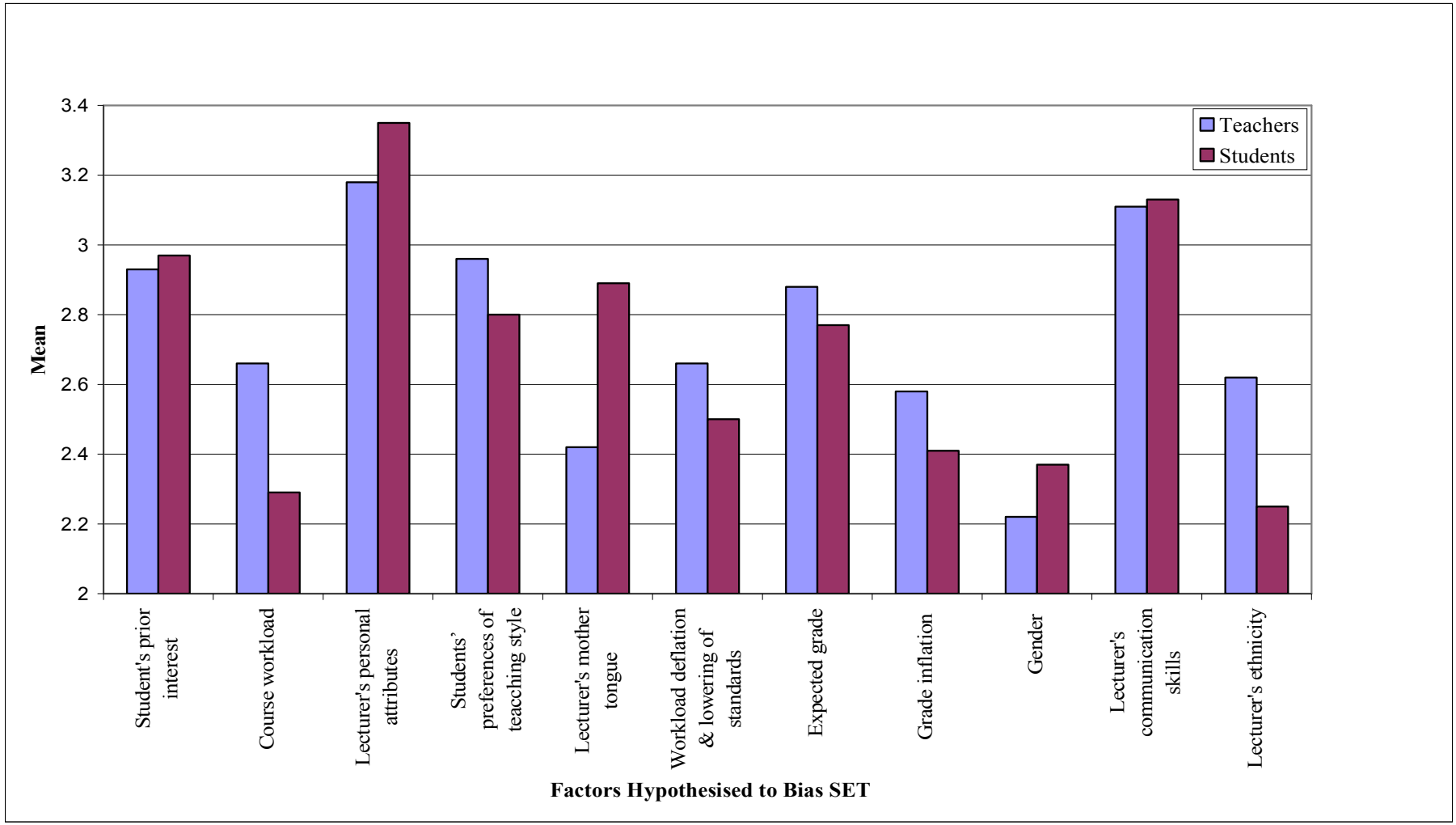
Teachers (*M*= 2.66, *SD*=.779) attached more weight than students did (*M*=2.29, *SD*=.898) to the effect of course workload on students' ratings, *t*= 6.494, *p* < .003. Teachers (*M*=2.62, *SD*=.894) were also more agreeable than students (*M*=2.25, *SD*=1.100) that students' evaluations of teaching could be affected by the teacher's ethnicity, *t*= 5.584, *p* < .003.

Students, however, seemed to agree more than the teachers did on the effect on students' ratings of two other factors. Students (*M*=3.35, *SD*=.744) showed stronger agreement than their teachers did (*M*=3.18, *SD*=.660) that lecturer's personal attributes could affect students' ratings, *t*= -3.471, *p* < .003. Moreover, students (*M*=2.89, *SD*=.909) placed more emphasis on lecturer's mother tongue as a biasing factor to students' ratings compared to the teachers (*M*=2.42, *SD*=.913), *t*= -7.087, *p* < .003.

On the other hand, teachers and students seemed to match closely in their assessment of the effect of the other seven hypothesized factors, namely: student's prior interest, students' preferences of teaching style, workload deflation and lowering of standards, expected grade, grade inflation, gender, and lecturer's communication skills.

As visually represented in Figure 7.1, two of the 11 hypothesized factors emerge as having the strongest effect on the validity of SET from both teachers' and student's points of view. These are lecturer's personal attributes and lecturer's communication skills. For both factors, students' assessment of the effect on the validity of SET is higher than that of the teachers, although the difference between the two is

Figure 7.1: Mean Teachers' and Students' Perceptions of the Effect of Various Factors Hypothesised to Bias SET



statistically significant only for the lecturer's personal attribute factor. This finding partially mirrors the finding in chapter six regarding the rating of teacher's friendliness and expressiveness as important traits of good college teaching. Both traits were rated higher by the students than by the teachers, but again with the statistical significance being in the difference over the personality aspect, i.e. being friendly.

These findings also point to a gap between teachers and their students in their understanding of the importance of lecturer's friendliness and expressiveness in effective college teaching and the possible effects or biases that lecturer's personal attributes may prompt in student evaluation of teaching. The findings also bring to light the issue of educational seduction, or Dr. Fox effect, discussed in chapter 4. Several researchers (e.g. Emery *et al.*, 2003; Naftulin, *et al.*, 1973) argued that students' ratings are more strongly influenced by the teacher's expressiveness, style, and personal attributes than by content. In a meta-analysis by Abrami, Leventhal, and Perry (1982) the researchers also concluded that expressiveness manipulations had substantial effects on students' overall ratings of their teachers, but limited impact on students' achievement. Content manipulations, on the other hand, were found to have substantial impact on achievement, but modest effects on students' evaluation of their teachers. These findings, however, were later contested by other researchers (e.g. Marsh and Ware, 1982), who argued that teacher's expressiveness affected only the ratings of the teacher's enthusiasm, which is the aspect of teaching most logically related to expressiveness manipulation.

7.1.2 Teachers' and Students' Perceptions of the Utility of Students' Evaluations of College Teaching

As discussed in chapter 4, the utility of SET is another area of heated debate in the literature about the evaluation of college teaching. While most researchers in the field agree that students' ratings of college teachers is the most widely used technique in measuring teaching effectiveness in higher education, there is a degree of disagreement over the utility of data obtained from these ratings. Evidence from the literature in the subject also suggests that data collected from students' ratings is more commonly used by administrators and faculty, as opposed to students themselves - in course and instructor selection, for example. Program administrators and faculty alike use SET data for various formative evaluation purposes, such as providing/obtaining diagnostic feedback on teaching performance and planning and directing teacher development and improvement efforts. However, some administrators also use data from student evaluation of teaching for summative evaluation purposes, such as assessing teacher performance for personnel decision-making purposes and in promoting accountability and quality control within their educational institutions. While this is probably the most disputed use of SET, the findings of the research literature on the effects of using students' ratings for making personnel decisions concerning faculty tenure and contracts, promotion, and salaries, are conflicting and inconclusive. Both the formative and summative approaches to SET continue to be used in higher education institutions today with varying degrees of success and sometimes conflict.

To explore research participants' views and opinions about the utility of students' ratings and how data from students' ratings can be used, four items were included in section two of the "Perceptions" questionnaire that focused on this aspect of SET.

Teachers and students were asked to indicate their level of agreement/disagreement with each use of SET using the scale provided. Table 7.3 below presents these four items and their abbreviations, which will appear in the tables, charts, and discussions that will follow.

Table 7.3: SET Uses and Their Abbreviations

	SET uses	Abbreviation
1.	Student ratings can provide reliable and valid diagnostic feedback to lecturers for improving teaching.	SET for diagnostic feedback & improvement
2.	Student ratings of lecturers should be used as one source of data for making personnel decisions such as lecturers' contract renewal or termination.	SET as a source of data for personnel decisions
3.	Students ratings of teaching can provide good protection to lecturers against potential biases in evaluation by heads of departments	SET for protection to lecturers against HoD's bias
4.	Student ratings of lecturers should only be used for teaching improvement purposes, not to hold lecturer accountable for deficiencies in their performance.	SET for improvement purposes only, not for accountability

Teachers' and students' perceptions of the utility of students' ratings of college lecturers were first compared by calculating the item mean for the 4 items listed above. A high mean value in this category of items would indicate agreement with the utility of SET suggested in each statement, while a low mean value would indicate disagreement with proposed usage of SET. An independent sample *t*-test was used to identify significant differences between the means of the two sample groups. These means and the t-test results for each item are compared for teachers and students in Table 7.4.

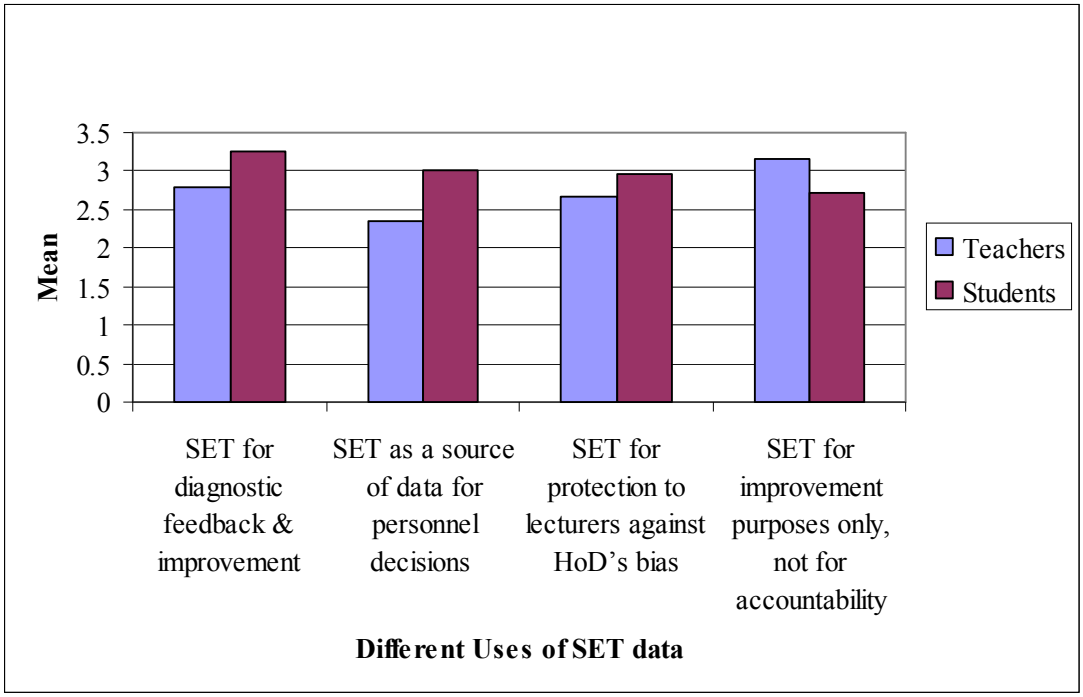
Table 7.4: Teachers' and Students' Perceptions of Various Uses of SET

	Teachers					Students						
	N	**Min**	**Max**	**Mean**	**SD**	**N**	**Min**	**Max**	**Mean**	**SD**	***t***	**Sig.**
SET for diagnostic feedback & improvement	248	1	4	2.79	.700	965	1	4	3.26	.664	-9.873	**.000**
SET as a source of data for personnel decisions	246	1	4	2.36	.858	960	1	4	3.02	.859	-10.809	**.000**
SET for protection to lecturers against HoD's bias	245	1	4	2.67	.796	952	1	4	2.95	.790	-5.030	**.000**
SET for improvement purposes only, not for accountability	244	1	4	3.15	.800	961	1	4	2.71	1.036	7.178	**.000**
*** Significant differences at $p < .003$ (2-tailed)**												

It can be seen from Table 7.4 and Figure 7.2 that teachers are generally less enthusiastic for SET than students. The t-test results also show that significant differences exist between the two sample groups in their views about the utility of SET.

The biggest difference (t= -10.809, p < .003) between teachers and students is on the use of SET as a source of data in making personnel decisions. Students (M= 3.02, SD= .859) are more enthusiastic than teachers (M= 2.36, SD= .858) for the use of data from students' ratings in making administrative decisions regarding teachers' tenure, promotion, and contract termination.

Figure 7.2: Mean Teachers' and Students' Perceptions of the Different Uses of SET



Perhaps not surprisingly, a number of other studies reached this same finding in the past. For example, in a survey by Sojka *et al.* (2002) the researchers found that 24% of the students in their sample believed that more weight should be given to student evaluations in promotion, tenure, and salary raise decisions. This is compared to 14.8% of faculty only who held the same view.

The second biggest difference (t= -9.873, p < .003) between teachers and students concerns the use of SET for giving diagnostic feedback to teachers on their teaching performance. Again, students (M= 3.26, SD= .664) seem to agree more than teachers (M= 2.79, SD= .700) do with using the results of students' ratings in providing diagnostic feedback to lecturers on their teaching performance in the classroom and in planning improvement efforts. This seems to contradict some of the findings of previous research where SETs were seen by teachers as a cause of improvement in

teaching (e.g. Murray, 1987; Sojka *et al.*, 2002). Summarising the results of published surveys from seven universities, Murray found that 80% of the teachers thought that SETs led to improvement in teaching. Sojka *et al.* (2002) also found that 44% of surveyed faculty (and only 24% of students) believe that teachers make significant changes in teaching style based on student evaluations.

Teachers' and students' perceptions are also significantly different (t= 7.178, p < .003) on whether the use of SET data should be limited to teaching improvement purposes only, and not for holding lecturers accountable for shortcomings in their teaching. Teachers (M= 3.15, SD= .800) expressed far stronger support than students did (M= 2.71, SD= 1.036) for limiting the use of results from students' ratings to improvement purposes only.

Finally, teachers and students also differ (t= -5.030, p < .003) on whether students' ratings provide protection to lecturers against biased evaluation by heads of departments. Students (M= 2.95, SD= .790) expressed stronger recognition of the usefulness of SET in this regard compared to the teachers (M= 2.67, SD= .796).

It is obvious from the findings presented and discussed above that the two sample groups – teachers and students- almost seem to "distrust" each other, especially on whether SET data should be used for making personnel decisions or just for improvement purposes. This is exactly what was found by Sojka *et al.* (2002).

> …faculty are very sensitive to the factors that are considered for tenure and promotion. Students, though, may be totally unaware of the politics of teaching and consequently fail to realize how evaluations may be used by administrators. Furthermore, suppose students did understand the effect that SET can have on faculty member's career. Would their responses be any different?
>
> (p.47)

This same question, whether the purpose of students' ratings affected their pattern, was investigated by Young *et al.* (1999). They concluded that the purpose of students' ratings (formative or summative) did not affect the pattern of these ratings. "That is, evaluations provided by students were similar for general directions lacking a purpose and for specific directions containing formative (instruction improvement) and summative (salary increase) purposes" (p.189).

7.1.3 Teachers' and Students' Perceptions of the Role and Involvement of the Student as an Evaluator of College Teaching

Although SETs are widely endorsed by teachers, students, and administrators (Marsh, 2007), a number of professionals and researchers cited in chapter four (e.g. Emery *et al.*, 2003; Naftulin *et al.*, 1973) question the ability of students to evaluate a multidimensional and complex activity like college teaching. As discussed in chapter two, however, a number of researchers (e.g. Boyer, 1990; Cashin, 1996; Kwan 2000; Loder, 1990; Marincovich, 1999; McKeachie, 1997b; Scriven, 1981) point out that students can be prepared for their role in SETs as evaluators and observers of teaching. They argue that there is a real need for colleges to involve students in constructing rating forms and to train them in using these forms by explaining to them how they have been constructed and what each item measures. It is believed that this would ultimately enhance the validity and reliability of SETs.

To explore research participants' perceptions of the role and involvement of the student in SETs, four of the items in the second section of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire were designed to measure respondents' approval/disapproval of four aspects of student's

involvement as an evaluator and observer of college teaching. These are presented and abbreviated in Table 7.5.

Table 7.5: Student's Role and Involvement in SET

Aspects of student's role and involvement in SET	Abbreviation
Students in the Foundation Program are capable of rating most aspects of a lecturer's teaching performance	Students are capable of rating teachers
Involving students and lecturers in developing rating forms will create a common ground for both parties to develop shared meanings of 'good' college teaching	Involving students and lecturers in developing SET forms
Students should be trained in using rating forms and told what each item represents before they are asked to rate their lecturers	Training students as evaluators
Educating students about the generic characteristics of effective college teaching will improve the reliability and validity of their ratings	Educating students about teaching effectiveness

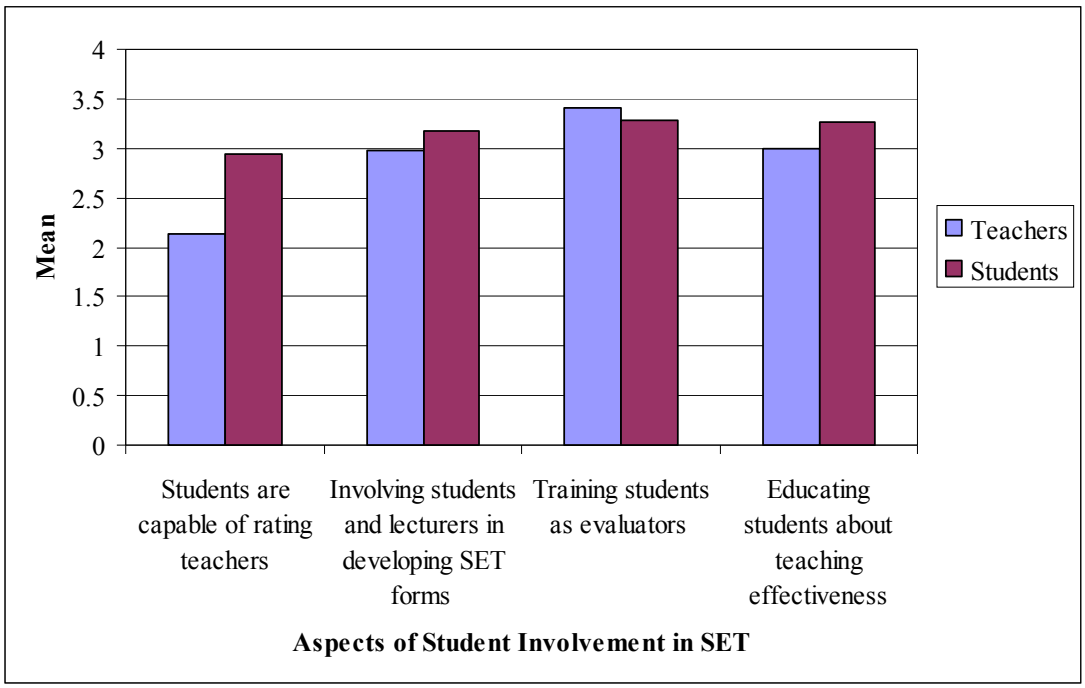
Again, the item mean was used to compare teachers' and students' opinions and perceptions of this aspect of SET. Independent sample t-test was used to identify significant differences between means (Table 7.6).

Table 7.6: Teachers' and Students' Perceptions of Student's Role and Involvement in SET

	Teachers					Students						
	N	Min	Max	Mean	SD	N	Min	Max	Mean	SD	*t*	Sig.
Students are capable of rating teachers	243	1	4	2.14	0.833	961	1	4	2.95	0.831	-13.496	**.000**
Involving students and lecturers in developing SET forms	239	1	4	2.98	0.730	960	1	4	3.18	0.748	-3.772	**.000**
Training students as evaluators	245	1	4	3.41	0.687	958	1	4	3.28	0.705	2.620	.009
Educating students about teaching effectiveness	237	1	4	3.00	0.698	963	1	4	3.26	0.665	-5.281	**.000**
*** Significant differences at $p < .003$ (2-tailed)**												

As represented graphically in Figure 7.3, a stronger role and involvement for students in SET is understandably more appealing to students than to lecturers. With the exception of "training students as evaluators", students' support for the proposed areas of student involvement in SET is significantly stronger than the lecturers'.

Figure 7.3: Mean Teachers' and Students' Perceptions of Various Aspects of Student Involvement in SET



The area where the biggest difference (t= -13.496, p < .003) between lecturers and students occurs is the area concerning students' capability to rate their lecturers' teaching performance. Students (M= 2.95, SD= .831) expressed stronger confidence in their ability to observe and rate most aspects of their lecturers' teaching performance compared the lecturers (M= 2.14, SD= .833). Students (M= 3.18, SD= .748) were also more enthusiastic than lecturers (M= 2.98, SD=.730) for the idea of involving lecturers and students in developing rating forms in order to develop shared meanings of effective teaching, t= -3.772, p < .003. Roche and Marsh (2002) argue that students' ratings in their own right enhance shared meanings of effective

teaching. This is because students' evaluations of their teachers' performance in the classroom make teachers' perceptions of their own teaching become more consistent with their students' perceptions. Richardson (2005) adds that "students' evaluations may change teachers' self-perceptions even if they do not change their teaching behaviour" (p.389).

As mentioned earlier, however, the idea of training students to use rating forms and explaining what each item in these forms mean sounds more appealing to teachers (M=3.41, SD= .687) than to students (M= 3.28, SD= .705). The difference between the two groups over this form of involvement, however, does not reach a statistical significance. Although less enthusiastic for being formally trained on using rating forms, students seem to show great interest in learning about the generic characteristics of effective college teaching. They also seem to recognise the positive effect this type of induction may have on the reliability and validity of their ratings, showing stronger agreement with it (M= 3.26, SD= .665) compared with their teachers (M= 3.00, SD= .698), t= -5.281, $p < .003$.

7.2 Chapter Summary

As stressed in the introduction of this chapter, it is important to understand how lecturers' and students' perceive SETs in order to interpret and use the data generated from them sensibly and effectively. Moreover, a great deal of the credibility, legitimacy, and usefulness of any teaching appraisal scheme lies in the acceptance and recognition it gains from its stakeholders and the shared meanings and understanding these stakeholders have for its purpose. Using the data derived from section two of the *Perceptions of Good College Teaching and Students' Evaluation of Teaching* questionnaire, the chapter compared and contrasted teachers' and

students' perceptions of SET in three specific areas: the factors hypothesised to bias students' ratings, the utility of students' ratings, and the role of the student in the evaluation of teaching.

Teachers significantly differed from students in their perception of the effect of four hypothesised biasing factors on the validity of SET, namely: course workload, lecturer's personal attributes, lecturer's mother tongue, and lecturer's ethnic background. The teachers were more agreeable than the students that course workload and teacher's ethnicity had an effect on students' ratings. Students, however, seemed to place more weight than the teachers did on the effect of lecturer's personal attributes and lecturer's mother tongue on students' ratings. Two of the 11 hypothesized factors, specifically lecturer's personal attributes and lecturer's communication skills, appeared to have the strongest effect on the validity of SET from both teachers' and student's points of view, but with students giving slightly greater emphasis to both factors.

Teachers and students did not seem to differ significantly in their assessment of the effect of the other seven hypothesized factors, namely: student's prior interest, students' preferences of teaching style, workload deflation and lowering of standards, expected grade, grade inflation, gender, and lecturer's communication skills.

With regard to the utility of SETs, in general, teachers were found to be less enthusiastic for SET than students were. Significant differences exist between the two sample groups in their perceptions of the utility of SET. The biggest difference

between teachers and students was on using students' ratings as a source of data in making personnel decisions. Consistent with the findings of previous research, students were found to be more enthusiastic than the teachers for using the results of students' ratings in making administrative decisions regarding teachers' tenure, promotion, and contract termination.

The second biggest difference recorded between the teachers and the students centres on the use of SET for giving diagnostic feedback to teachers to improve their teaching. Students seem to agree more than the teachers do with using the results of SET in providing diagnostic feedback to college teachers on their teaching effectiveness with the aim of improving performance. Teachers also expressed significantly stronger support for limiting the use of SET to improvement purposes only. Teachers and students also differed on whether students' ratings protected teachers against biases in administrator evaluations, with students expressing stronger recognition of this use compared to their teachers.

The role of the student in the evaluation of teaching itself was generally an area of disagreements between the teachers and the students. The biggest difference between the two groups was in the very heart of the argument surrounding SET- whether students are capable of evaluating or rating their teachers. Students expressed stronger confidence in their ability to evaluate most dimensions of their lecturers' teaching compared to the teachers. Students also showed more enthusiasm than the teachers for working together in developing rating instruments and for learning about the generic characteristics of effective college teaching and seemed to recognise the potential of this in improving the reliability and validity of their ratings more than

their teachers. However, students were found to be less enthusiastic than the teachers for receiving formal training on using rating forms. The difference between the two groups over this aspect, however, does not reach a statistical significance.

To sum up, teachers and students apparently have different perceptions of SET, their validity, utility, and the role and involvement of the student as an evaluator of teaching. In line with the findings of Ryan *et al.* (1980), two of the most significant differences between teachers and students in their perceptions of students' evaluation of college teaching seem to centre around two main issues: (a) students' ability and expertise to evaluate college teaching appropriately and accurately, and (b) the use of data from students' ratings in making personnel decisions, such as contract renewal, promotion and tenure.

CHAPTER EIGHT

THE MULTI-DIMENSIONALITY AND RELIABILITY OF STUDENTS' EVALUATIONS OF COLLEGE TEACHING: EVIDENCE FROM THE TRIAL OF *SEEQ* IN OMAN

8.0 Introduction

In chapters 6 and 7, teachers' and students' perceptions of the characteristics of effective college teaching as well as their views on the validity, utility, and the role of the student in SET were compared and discussed. In this chapter, the findings about students' ability to identify the various dimensions of teaching effectiveness underlying SEEQ-a widely used western standardised rating instrument- and their ability to give reliable ratings of classroom instruction using the same instrument will be examined and discussed. In addition, the effect of various student, lecturer, and course background characteristics that are traditionally hypothesised to bias students' ratings will be investigated.

As mentioned in Chapter 5, although the SEEQ is a widely used SET instrument, it has never been used in Oman before as a research tool or as a questionnaire for collecting students' ratings. Furthermore, there seems to be no research available on the applicability, factor stability, or reliability of this instrument in the Omani context or even in the Arab world. To this end, this chapter is set to answer three main research questions:

- **Research question 5:** What dimensions of teaching underlie students' evaluations of teaching in Oman and to what extent are these dimensions

similar to or different from the dimensions of teaching identified in the relevant western SET literature?

- **Research question 6:** How reliable are college students' evaluations of teaching in the Omani context?
- **Research question 7:** To what extent do student, lecturer, and course background characteristics influence students' ratings?

In order to answer theses three questions and evaluate the appropriateness of use of the SEEQ questionnaire in the Omani context, a factor analysis and a reliability analysis were carried out on the data collected from the research context. Furthermore, a number of tests were carried out on the ratings collected with SEEQ to determine the nature of influence and effect size of various background characteristics on students' evaluation of teaching effectiveness.

8.1 Research question 5: What dimensions of teaching underlie students' evaluations of teaching in Oman and to what extent are these dimensions similar to or different from the dimensions of teaching identified in the relevant western SET literature?

In line with the procedures followed in SEEQ research investigating the instrument's applicability across contexts, an exploratory factor analysis was carried out on the data gathered from the student sample using this instrument. Factor analysis is a 'data reduction' technique used to identify "'clumps' or groups among the intercorrelations of a set of variables" (Pallant, 2007: 179). In SEEQ research, one main application of this procedure is "to describe the different components of teaching effectiveness actually being measured by a set of questions" (Marsh, 1982a: 79). The factor analysis in the present study is designed to serve two main purposes

which are often highlighted in similar SEEQ factor analysis research (e.g. Marsh & Hocevar, 1991b). These two purposes are: a) to test whether students in the sample were able to identify the different components of effective teaching; b) to test whether the factors emerging from the data replicate the factors that the SEEQ instrument was designed to measure.

8.1.1 Establishing the Suitability of the Data for Factor Analysis

Prior to performing factor analysis, the suitability of the data set for factor analysis was assessed. Firstly, the sample size was checked against the parameters recommended in the research literature. Despite the lack of agreement among authors regarding the overall sample size needed for factor analysis, "the recommendation generally is: the larger, the better…[as] factors obtained from small data sets do not generalise as well as those derived from larger samples" (Pallant, 2007: 180-181). Some authors emphasise the ratio of subjects to items rather than the overall sample size. For example, Nunnally (1978) recommends ten cases for each item to be included in the factor analysis. Yet there are studies which reported factor analyses based on cases to item ratios of only 2:1 or even less (Costello & Osborne, 2005). Other studies, on the other hand, have shown that adequate sample size is partly determined by the nature of the data itself and is variant across studies (MacCallm, Widaman, Zhang, & Hong, 1999; Costello & Osborne, 2005). "In general, the stronger the data, the smaller the sample can be for an accurate analysis" (Costello & Osborne, 2005: 4). By strong data, Costello & Osborne mean high communalities without cross loadings. Bearing in mind the highly intensive multivariate technique involved in factor analysis, and because of the uncertainty in the factor analysis sampling theory and to minimise the effect of sampling error, the decision was made

in the present study to keep the subjects to item ratio to around 30:1 (around 30 cases for each of the 31 SEEQ items included in the factor analysis).

Secondly, the results of the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and the Bartlett's test of sphericity were examined to assess the factorability of the data. The KMO was found to be .96, exceeding the recommended minimum of 0.60 (Kaiser, 1974; Pallant, 2007; Tabachnick & Fidell, 2007). "This result indicates that the degree of common variance among the items is very good and that it was appropriate to apply factor analysis" (Penny, 2004: 185). This "superb" KMO value (Hutcheson & Sofroniou, 1999) also indicates that the sample size is adequate for factor analysis. The Bartlett's test of sphericity was also highly significant ($p < 0.001$). Inspection of the correlation matrix also revealed the presence of many coefficients of .3 and above. This suggests that "correlations are real and not attributable to chance or sampling error" (Penny, 2004: 185), supporting the factorability of the correlation matrix.

8.1.2 Exploratory Factor Analysis

Replicating similar procedures used by Marsh and Hocevar (1991b) and Penny (2004), the number of factors to be extracted was determined a priori. Therefore, the analysis did not rely solely on the conventional techniques of inspecting the scree plot and retaining the components with eigenvalues exceeding one. The 29 items composing the main eight sub-scales (factors) of SEEQ, in addition to the two Overall Rating items[4], were subjected to principal axis factoring followed by Varimax rotation using SPSS Version 15. These techniques of extraction and rotation

[4] These two items do not constitute a separate factor. Item 30 (overall course rating) is designed to load with the Learning/Academic Value factor, while item 31 (overall teacher l rating) is designed to load with Lecturer Enthusiasm factor (Marsh & Hocevar, 1991b).

were found to yield optimum SEEQ factor solution, compared to principal components analysis and oblique rotation, which are also commonly used in SEEQ research (Penny, 2004).

> In principal axis factoring only common or shared variance is analysed, compared with the assumption of the principal components analysis extraction method, for example, that all the variability in an item should be used in the analysis. A varimax rotation attempts to achieve a 'simple structure' by minimising the number of items that load highly on a factor, using the orthogonal assumption that the factors are uncorrelated. The aim here is to make the factor solution more interpretable.
>
> Penny (2004: 183-184)

It is also argued that factor analysis, like principal axis factoring, is preferable and more superior to principal components analysis. The latter may be considered more of a data reduction method while the first is a technique used to uncover the constructs underlying the data (Ford, MacCallum, & Tait, 1986).

This 'forced' 8 factor solution produced five components with initial eigenvalues exceeding 1, explaining 38%, 6.1%, 4.1%, 3.6%, and 3.4% of the variance respectively. The other three factors, however, had initial eigenvalues of less than 1, explaining 2.9%, 2.6%, and 2.6% of the variance respectively. This brings the total variance explained by the eight factors together to 63.3%. An inspection of the scree plot revealed a clear break after the third factor and another much smaller break after the fifth factor.

Separate trial runs of factor analysis forced to seven, six, and five factors were carried out for exploratory purposes and to compare the interpretability of the resulting factor structures. However, it was noticed that the smaller the number of factors selected, the more item cross-loadings and/or factor overlaps are produced, despite experimenting with various extraction and rotation techniques. In addition, it

was evident that the smaller the number of factors retained, the lower the communalities became, which represented an unnecessary loss of information. Bearing in mind the exploratory nature of the analysis, it was found necessary that the factors retained should map as closely as possible onto the original variables. As Field (2009) argues:

> If the communalities represent a loss of information then they are important statistics. The closer the communalities are to 1, the better our factors are at explaining the original data. It is logical that the more factors retained, the greater the communalities will be (because less information is discarded); therefore, the communalities are good indices of whether too few factors have been retained.
>
> (pp. 641-642)

With an interpretable factor solution seeming more difficult to obtain using less than 8 factors, it was found necessary to retain the more interpretable 8 factor structure. The resulting factor structure is presented in Table 8.1. In addition to the item loadings for each factor, the communality value, h^2, is shown in the last column. Communality is the proportion of common variance present in a variable and shared with other variables in the common factor. None of the items has a communality value of less than .33, which is an indication that all items fit well, in a low to moderate degree, with the other items in their factors. High item communalities of .8 or greater are unlikely in social sciences, where low to moderate communalities of .40 to .70 are more common (Costello & Osborne, 2005). The factor structure also indicates satisfactory loadings, .34 to .73, for the items on the targeted factors, with median loading of .57. Tabachnick and Fidell (2001) cite .32 as the minimum loading of an item. Costello & Osborne (2005) also cite .32 as a cutting point in specifying item cross-loadings. "A "cross-loading" item is an item that loads at .32 or higher on two or more factors" (Costello & Osborne, 2005: 4). In light of these recommendations, and for ease of presentation, only item loadings of .32 or higher are displayed in Table 8.1.

It can be seen from Table 8.1 that all the items in the Learning/Academic Value, the Individual Rapport, and the Assignments/Readings scales loaded higher on the factors they are designed to measure with no cross-loadings at all. In the Lecturer Enthusiasm scale, item 8 (Presentation style holds your interest) loaded higher on the intended factor at .59, but cross-loaded with the Group Interaction factor items at .39. In the Group Interaction scale itself, all the items loaded higher on the targeted factor with no cross-loadings, except for item 13 (Students encouraged to participate). This item, although loading higher on the targeted factor at .56, cross-loaded at .36 with the Lecturer Enthusiasm scale.

The underlying construct of the Organisation/Clarity scale seems to overlap partially with both, the Lecturer Enthusiasm and the Group Interaction scales. In this scale, item 10, "Materials well prepared and explained", and item 11, "Lessons agreed with course objectives", loaded higher on the targeted factor, with the latter also cross-loading with the Group Interaction items at .33. The other two items, "Explanations are clear" and "Lectures facilitate note taking" loaded higher on the Lecturer Enthusiasm factor instead (represented in italics) and also cross-loaded with the Group Interaction factor.

Table 8.1: Factor Analysis Results of the SEEQ Items for the Total Sample

SEEQ Scale/Item (Abbreviated)		Factor Loadings							
		i	**ii**	**iii**	**iv**	**v**	**vi**	**vii**	h^2
Learning/Academic Value									
1.	Course challenging and stimulating	**.54**							.33
2.	Learning something valuable	**.73**							.57
3.	Interest in subject increased	**.59**							.40
4.	Learning and understanding materials	**.44**							.34
Lecturer Enthusiasm									
5.	Enthusiastic about teaching		**.68**						.65
6.	Dynamic and energetic		**.69**						.67
7.	Enhances presentation with humour		**.34**						.45
8.	Presentation style holds your interest		**.59**		.39				.69
Organisation/Clarity									
9.	Explanations are clear		*.52*		.39				.56
10.	Materials well prepared and explained			**.52**					.53
11.	Lessons agreed with course objectives			**.47**	.33				.52
12.	Lectures facilitate note-taking		*.44*		.41				.50
Group Interaction									
13.	Students encouraged to participate		36		**.56**				.51
14.	Students invited to share ideas				**.60**				.46
15.	Students encouraged to ask questions				**.53**				.51
16.	Encouraged to express own ideas				**.57**				.51
Individual Rapport									
17.	Friendly toward individual students					**.53**			.38
18.	Welcomes students seeking help					**.71**			.66
19.	Genuine interest in individual students					**.34**			.41
20.	Accessible to students during/after class					**.34**			.35
Breadth of Coverage									
21.	Contrasts various theories/concepts				*.44*				.46
22.	Presents background of concepts		.35		*.43*				.52
23.	Presents different points of view				*.44*				.37
24.	Discusses developments in the field				.36			*.37*	.50
Assessment/Grading									
25.	Feedback on assessments valuable						**.40**	.32	.52
26.	Evaluation methods fair/appropriate						**.48**		.45
27.	Assessments test course content						.36	*.38*	.42
Assignments/Readings									
28.	Required readings/texts valuable							**.67**	.59
29.	Assignments enhance understanding							**.71**	.67
Overall Rating									
30.	Overall course rating		*.58*						.58
31.	Overall teacher rating		**.65**						.66

Notes: h2 = Communality. N= 46 classes (922 students). Extraction method: Principal Axis Factoring. Rotation method: Varimax with Kaiser normalisation. Factor loading in **bold** are for items designed to measure each factor.

Cross-loadings can also be found in the Assessment/Grading scale. While item 26, "Evaluation methods fair/appropriate", loaded exclusively on its target factor at .48, item 25, "Feedback on assessments valuable", loaded higher on its factor but also cross-loaded at .32 with the Assignments/Readings items. Item 27, "Assessments test course content", not only cross-loaded with the Assignments/Readings factor, but also loaded higher (.38) on this factor as opposed to the loading on it own factor (.36). As for the Overall Rating items, it is worth repeating here that the Overall Rating items are not designed to rate a specific dimension of teaching effectiveness like the other eight scales. Rather, they are linked to the Instructor Enthusiasm and the Learning/ Academic Value factors (Marsh & Hocevar, 1991b). In this present study, both of the Overall Rating items loaded higher under the Lecturer Enthusiasm factor.

Probably the least defined factor resulting from this analysis is the Breadth of Coverage. None of the items loaded on its target factor. Instead, three of its items, 21, 22, and 23, loaded under the Group Interaction scale at .44, .43, and .44 respectively. Item 22, "Presents background of concepts" also cross-loaded with the Lecturer Enthusiasm factor at .35. The last item in this scale, "Discusses developments in the field" cross-loaded under Assignments/Readings and Group Interaction scales at .37 and .36 respectively.

Clearly, the SEEQ factor structure that Marsh found in America (Marsh, 1982a) and Australia (Marsh, 1981), and that was found by him and/or others in Spain (Marsh, Touron, & Wheeler, 1985) , New Zealand (Watkins, Marsh, & Young, 1987), and the UK (Coffey & Gibbs, 2001) was not duplicated precisely in Oman. Of the eight factors included in the analysis, only two factors, namely: Learning/Academic Value

and Individual Rapport, loaded independently and did not overlap with any other factors or had any cross-loading items. Three other scales did load higher under the targeted factors, but had a cross-loading item (Lecturer Enthusiasm and Group Interaction) or overlapped with non-target cross-loadings from other factors (Assignments/Readings). Nevertheless, the above indicated five factors can be clearly identified with the factor structure found in the western SEEQ research despite the few cross-loadings.

Examining the remaining three factors, Organisation/Clarity, Breadth of Coverage, and Assessment/Grading, however, one can observe important differences between Omani students and Western students. Omani students tend to link lecturer's personal and teaching/presentation skills (i.e. Enthusiasm and Group Interaction factors) with his/her organisational and planning capacity (i.e. Organisation/Clarity factor), hence the multiple cross-loadings and overlaps in the Organisation/Clarity scale with the other two factors shown in Table 8.1. Students in the present study also could not distinguish Breadth of Coverage as an independent dimension of effective teaching. Instead, they viewed most of its items as part of the Group Interaction factor.

The overlap between some factors found in this study replicates similar findings reached by a number of researchers who investigated the applicability of SEEQ in various developing countries (e.g. Clarkson, 1984; Lin, Watkins, & Meng, 1995; Watkins & Akande, 1992; Watkins & Regmi, 1992; Watkins & Thomas, 1991). Clarkson (1984) tested the applicability of SEEQ with Papua New Guinean university students and concluded that students from the developing country do not distinguish between the organisational skills and techniques of the lecturer and the rapport

created between the lecturer and students. Watkins & Akande (1992), Watkins & Regmi, (1992), and Watkins & Thomas (1991) also investigated the cross-cultural validity of SEEQ's multi-dimensional model of teaching effectiveness in Nigeria, Nepal, and India respectively and concluded that a significant degree of overlap existed between aspects of teaching skill and teacher enthusiasm in the SEEQ factor structure obtained from college students in these three countries. A similar study was also carried out by Lin, Watkins, and Meng (1995) which indicated more overlap among the dimensions of Learning/Value, Organisation/Clarity, Enthusiasm, and Breadth of Coverage, than that found in Western studies.

The other thing to note here, however, is that GFP students in Oman were able to isolate multiple factors despite the overlapping and cross-loadings between some factors. This indicates that Omani college students can identify different dimensions of a lecturer's teaching performance.

The Organisation/Clarity scale's overlapping with the Lecturer Enthusiasm and Group Interaction scales could be attributed to a number of factors. Students probably were unable to distinguish between the items rating their lecturer's clarity of explanations (item 9), success in defining and meeting the course objectives (item 11), and delivery of well-organised lectures that facilitated note-taking (item 12) and those items rating his/her enthusiasm (items 5-8) and management of group interaction (items 13-16). Students seem to expect a teacher who is enthusiastic about the subject and teaching to be well prepared for his/her classes and to give clear explanations and well-organised lectures/tutorials. A teacher who is well-prepared, well-organised, and energetic in the classroom and who makes judicious use of

humour in teaching and impresses students with his/her presentation styles is also more likely to have students who are motivated and interested in the subject. This motivational potential of classroom social interaction is what the Group Interaction factor is designed to measure. In this sense, the three factors are interrelated. This interrelation is even stronger in the context of the study where the teacher traditionally takes a centre stage in the teaching/learning activities and where teacher dependence is deeply rooted and prevalent in the culture of the educational organisation.

As mentioned before, the Breadth of Coverage items, unlike the other seven scales, did not yield any loadings on the targeted factor. Instead, three of the items, 21, 22, and 23, loaded under the Group Interaction factor and the remaining item, 24, loaded higher under the Assignments/Readings factor. It is not clear why GFP students' ratings in the present study completely failed to detect the construct underlying this factor. It can be argued that this scale is best suited to rating content-based classes rather than TESOL classes, which are mostly skill-based and primarily concerned with improving students' English language proficiency and providing opportunities for language acquisition and practice rather than delivery of content.

The partial overlapping between the Assessment/Grading scale and the Assignments/Readings scale is interesting. Assignments, readings, and homework, although assessed in the GFP, are not credited and do not contribute to the final mark in the GFP courses sampled in this study. This is probably why two items in the Assessment/Grading scale, "Feedback on assessments valuable" and "Assessments test course content" cross-loaded with the Assignments/Readings items.

In summary, the factor solution of the SEEQ obtained from the data in this study partially replicates the factor structure of this instrument reported in the research literature, but with low to moderate loadings on the targeted factors. Nine items had cross-factor loadings of .32 or higher under two factors. Partial overlapping is observed between the Organization/Clarity, the Lecturer Enthusiasm, and the Group Interaction scales. The factor structure of the Breadth of Coverage scale is probably the least interpretable in the whole matrix. None of the items in this scale loaded under the target factor. Instead they loaded and cross-loaded under three different and seemingly unrelated factors, namely Group Interaction, Lecturer Enthusiasm, and Assignments/Readings. Students also seem to have mixed certain aspects of Assessment and Grading with Assignments and Readings. From the above, it can be concluded that although Omani students can distinguish certain dimensions of effective teaching identified in the western SET research, some factors of effective teaching that underlie SET instruments in Western universities and colleges may not be separable in Oman. Moreover, further investigation on the applicability of the Breadth of Coverage scale to TESOL programs is required if SEEQ is to be used for collecting students' ratings in the GFPs.

8.2 Research question 6: How reliable are college students' evaluations of teaching in the Omani context?

As discussed in Chapter 4, many teachers as well as program administrators are suspicious about the reliability of students' ratings as a source of information regarding teaching effectiveness despite the large body of literature supporting the inter-rater reliability and consistency of SET. In the context of the study, these suspicions have neither been supported nor challenged by research findings so far.

In line with the research in the field (e.g. Marsh & Dunkin, 1992; Marsh & Roche, 1993; Cashin, 1995; Marsh, 2007, Penny, 2004), inter-rator reliability of the students' responses in the present study was estimated from the *class-average* response using the intraclass correlation coefficient (ICC). The analysis was carried out using MLwiN 2.17, SPSS Version 15, and Excel 2003. The ICC is obtained with a one-way analysis of variance (ANOVA) that compares the within-group variability with the between-group variability. As responses from students in a class ought to be more alike than students from different classes, because of their common experience, estimates of reliability of class-average responses are therefore expected to be high when differences between classes are much larger than differences within classes. It is important to note here that the reliability of the class-average response depends upon the number of students rating the class. It is estimated that the ICC reliability would be around .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for five students (Marsh, 2007).

Unlike other similar studies where the range in class size was big, the difference between the minimum class size and the maximum class size in the present study (minimum= 8; maximum= 24; mean= 20) was indeed very small to warrant separate ICC analyses to determine the effect of class size on intraclass correlation coefficient. Only one class had less than 16 students and the rest had between 16 and 24 students. Nevertheless, the effect of two other class characteristics, namely GFP level and course type, on the inter-rater reliability of students' ratings was examined. The decision to examine the potential effect of these two factors on the inter-rater reliability of students' ratings was primarily driven by the findings of the exploratory study. In the exploratory study, the GFP administrators and lecturers perceived

students' "lack of experience" and students' inability to judge certain "trade secrets" or teaching techniques in some areas of language classes as serious threats to students' ability to make reliable judgments about the quality of teaching they receive. The analysis was run for the total scale and for sub-scale scores, for the total sample of classes and then for different GFP levels and course types sampled in the study. A total of 46 classes (922 students) were included in the analysis. The results of the analysis are presented in Table 8.2.

8.2.1 Inter-rater Reliability Estimates for the Total Sample

Based on the results from the total sample of 46 classes (922 students), the median inter-rater reliability coefficient for the eight sub-scales and the two overall rating items is r=.87 for an average class size of 20 students (Table 8.2). Although this reliability coefficient is lower than the r=.89 found in Australia by Marsh & Roche (1993) and the r= .90 found in North America by (Marsh, 1987), the results from SEEQ trial in Oman in general do provide evidence of reasonably high inter-rater reliability for GFP students' ratings. This is despite the fact that, unlike North America, where SEEQ was first developed, and Australia, where the instrument has been widely used in research and teaching evaluation, SEEQ has never been used in Oman before. Furthermore, the majority of the students involved in this study have never been asked to systematically rate their teachers using a SET form before, the thing that adds to the unfamiliarity of the task to the students- as a practice, but not necessarily as a concept.

Table 8.2: SEEQ Sub-Scale Inter-rater Reliability Estimates For Total Sample and For Classes Differing In GFP Level and Course Type

SEEQ Sub-scales	Total sample	Class GFP Level			Course Type			
		Elementary Level	Intermediate Level	Advanced Level	Writing Skills	Reading Skills	Listening & Speaking Skills	Core Course (integrated skills)
Learning/ Academic Value	.77	.81	.68	.77	.72	.85	.54	.66
Lecturer Enthusiasm	.93	.84	.94	.94	.90	.91	.87	.95
Organization/ Clarity	.92	.81	.95	.91	.92	.92	.61	.95
Group Interaction	.87	.67	.89	.89	.89	.87	.65	.88
Individual Rapport	.85	.62	.83	.89	.76	.85	.47	.91
Breadth of Coverage	.87	.79	.90	.87	.87	.87	.44	.91
Assessment/Grading	.87	.84	.87	.86	.77	.88	.64	.90
Assignments/Readings	.86	.91	.86	.76	.81	.85	.72	.89
Overall rating item (course)	.87	.82	.84	.90	.89	.90	.77	.79
Overall rating item (Lecturer)	.91	.84	.93	.92	.89	.91	.82	.93
Total scale	.93	.89	.94	.93	.92	.94	.75	.95
Summary Statistics								
No. of Classes	46	10	18	18	11	10	13	12
No. of Students	922	185	362	375	211	210	257	244
Minimum Class Size	8	8	16	17	16	17	8	16
Maximum Class Size	24	24	23	24	23	24	24	23
Mean Class Size	20	18.5	20.1	21.8	19.2	21	19.8	20.3

For the total sample of 46 classes, the sub-scale inter-rater reliability estimates were generally high. Apart from a moderate r=.77 for the Learning/Academic Value scale, none of the other scales or overall rating items yielded an inter-rater reliability of less than r=.85. The highest value, r=.93, was for the Lecturer Enthusiasm scale. The moderate coefficient for the Learning/Academic Value scale can be explained, at least in part, by the fact that the items in this scale do not rate observable teaching behaviours. Rather, the scale measures students' satisfaction with their own learning experience or gain in the course. As discussed above, however, students in the context of the present study are generally accustomed to classes in which the teacher and what he/she does in the class is the focus of attention, not what students do or learn. Therefore, it may be much easier for the students to evaluate the observable teaching behaviour of their instructor than to assess their own learning from the course.

8.2.2 Inter-rater Reliability Estimates for Classes Differing in GFP Level

When examining the inter-rater reliability for the different GFP Levels, the median for the eight sub-scales and the two overall rating items was also found to be good (Table 8.2). It was r=.82 for Elementary, r=.88 for Intermediate, and r=.89 for Advanced. It is clear that the inter-rater reliability increases with the Level. The difference between the median reliability estimate for the Elementary Level (.82) on one side and the median reliability estimates for the Intermediate and Advanced Levels (.88 and .89 respectively) on the other side is particularly large. While the inter-rater reliability estimates are known to vary systematically with class size (Marsh & Roche, 1993), this observed difference in the median reliability of ratings between the Elementary Level (average class size of 18.5) and the Intermediate and Advanced Levels (average class size of 20.1

and 21.8 respectively) could not be attributed only to the difference in class size. According to Marsh & Roche (1993), the median estimates for groups of classes with an average class size of 16 students and 21 students are .83 and .86 respectively. In the present study, the difference in the average class size between the Elementary Level and the Advanced Level is marginal- around three students only- and, therefore, is not sufficient to explain the wide gap in reliability estimates between the two Levels.

This gap can be attributed, in part at least, to students' increased acquaintance with the characteristics of college teaching as they progress through the different levels. The improvement in students' linguistic ability in the target language as they move form the entry level to the exit level of the GFP could also have a role in facilitating communication between students and their teachers. The language barrier could be a source of frustration, misunderstanding, and distrust between students and their teachers in lower GFP Levels, the thing which may reflect negatively on students' ability to make reliable evaluations of their teachers. In colleges where students' ratings are collected, students' familiarity with the practice of SETs and their perceived benefits and implications may also help improve students' understanding of their role as college students and as evaluators of teaching. Ideally, the further the students progress into a program of study, the more aware they become of its requirements, distinctive teaching styles, and assessment methods.

As is apparent from the table, the sub-scale inter-rater reliability estimates across the different levels are also generally high and above the acceptable level of r=.70, with few exceptions. For the Elementary Level, the reliability estimates ranged from r=.62 for the

Individual Rapport scale to *r*=.91 for the Assignments/Readings scale. In this Level, however, two scales resulted in relatively low inter-rater reliability estimates, namely the Group Interaction scale (*r*= .67) and the Individual Rapport scale (*r*=.62). There could be a number of reasons behind this low reliability coefficient. Elementary Level is the entry level in the GFP for students who fail to secure the required scores in the Placement Test for entry into the higher Intermediate and Advanced Levels. A student is given the chance to repeat one level only and in case he/she fails again in the same level or in a subsequent level, that student is expelled from the GFP program and the college. Of course, the more levels for a student to go through, the bigger the chance that he/she may exhaust the one-time-repeat chance before reaching the exit level. Bearing in mind their modest English proficiency level and the pressure these freshmen feel in coping with the demands of the system and the new teaching/learning environment and medium of instruction, students in this level usually experience a tremendous level of anxiety and frustration.

This frustration usually surfaces in situations which are in sharp contrast to what students are used to in pre-college education, particularly in the areas of group interaction and individual rapport between teachers and their students. Students tend to confuse a lecturer's need to conform to the system's rules and regulations with arrogance and lack of cooperation and respect. As discussed earlier in Chapter 3, showing respect for students is the single most important characteristic of a good teacher in the eyes of Arab students in the Gulf (Saafin, 2008). As evident from Saafin's study, however, Arab Gulf students also consider their teachers' "willingness to compromise" as the second most important trait of an effective teacher. While this definition of

effectiveness may be tolerated in schools which are primarily staffed by indigenous teachers or teachers form neighbouring Arab countries with a similar culture, it could be extremely difficult for a multi-national faculty with multi-cultural backgrounds to embrace this understanding of "effectiveness". The result may be continuous conflicts and disagreement between students and their teachers over what constitutes good rapport and healthy group interaction.

The differences in inter-rater reliability between the Intermediate and Advanced Level groups are minor. In the Intermediate Level, the reliability estimates ranged from a relatively low r=.68 for the Learning/Academic Value scale to a very high r=.95 for the Organisation/ Clarity scale. Inter-rater reliability estimates for SEEQ scales at the Advanced Level were also high in general, ranging from r=.76 for the Assignments/Readings scale to r=.94 for the Lecturer Enthusiasm scale. As can be seen in Table 8.2, the inter-rater reliability estimates for the Learning/ Academic Value scale in these two Levels are again within the low to moderate range, probably for the same reasons identified earlier for the total sample. One significant improvement in these two Levels is the noticeable increase in the reliability estimates for the Group Interaction and Individual Rapport scales compared to the Elementary Level. It can be assumed here that as students progress through the GFP Levels, their understanding of the system, course requirements, and the multi-cultural backgrounds of the teaching staff improve and, consequently, their ability to make reliable judgments about the effectiveness of their lecturers in managing group interaction and individual rapport with students also improves. Where SET ratings are collected, the practice of giving feedback on the quality of teaching itself may help alleviate some of the causes of misunderstanding or

mismatched expectations between teachers and their students over time. Roche and Marsh (2002) found that teachers' understanding of their own teaching effectiveness became more consistent with students' perceptions of their teaching as a result of receiving students' ratings. Also, as mentioned in chapter 7, students' evaluations may have a strong effect in changing teachers' self-perceptions, even if they do not directly induce change in teaching behaviour (Richardson, 2005).

However, this improvement seems to be coupled with a change in the reliability estimates of the Assignments/ Readings scale, but, surprisingly, in the opposite direction. The Advanced Level groups yielded significantly lower inter-rater reliability estimate (r=.76) on this scale compared to the Elementary (r=.91) and Intermediate (r=.86) Level groups. This probably is because assignments, homework, and extra readings in the GFP are not credited and do not contribute to the continuous assessment mark of any of the courses sampled in the study. As highlighted in a quality assurance manual used by one of the colleges involved in the study, the allocation of [continuous assessment] marks for each Skill is based on quizzes only. This is being the case, it can be said that many GFP students may gradually lose interest and faith in the value of homework and assignments and their contribution to the appreciation and understanding of the subject. As a result, the level of agreement between students on the value of homework and assignments diminishes over time.

8.2.3 Inter-rater Reliability Estimates for Classes Differing in Course Type

The median inter-rater reliability estimates for SEEQ's eight scales and the two overall rating items at the course-type level were also good, except for the Listening and Speaking Skills course (Table 8.2). They were r=.88 for Writing Skills, r=.88 for Reading Skills, r=.91 for the Core Course (integrated skills), but only r=.65 for Listening & Speaking Skills.

It is not clear why students' ratings in the Listening and Speaking Skills courses yielded lower inter-rater reliability estimates. These results are particularly surprising when we know that 75% of the students who rated their Listening and Speaking course lecturers are from the Intermediate and Advanced Levels, which demonstrated higher reliability estimates compared to the Elementary Level. One possible factor that may have negatively affected the reliability of students' evaluations of their Listening and Speaking course tutors can be traced back to students' perceptions of the importance of certain characteristics of college teachers discussed in Chapter 6. In both the exploratory study findings and the findings of the main study which in part investigated students' and lecturers' perceptions of effective teaching (Chapter 6), many students expressed preferences for lecturers who are native speakers of English and/or who use native-like intonation and stress. This may have lead some students to believe that being a native speaker of English per se is a prerequisite to teaching effectiveness in TESOL classes. In classes taught by non-native speakers of English, such beliefs may translate into polarised SET ratings.

Listening and speaking skills are probably the most used skills in everyday communication. In a second/foreign language class, listening and speaking may also be the most difficult language skills to develop. Under the increasing demand for competent users of English in the era of globalisation, policy makers as well as prospective employers in the labour market in Oman and elsewhere in the developing world have put educational institutions under great pressure to improve the quality of their English language teaching (ELT) programs. Teachers became under more pressure to prove their competence and skills in ELT to their direct employers and to the business community and industry. Developing students' communication skills in English became an area in which teachers were encouraged and sometimes required by their direct supervisors to demonstrate their up-to-date knowledge and application techniques of prevailing and widely accepted approaches in language teaching, most notably Communicative Language Teaching (CLT), without careful consideration to the context or students' preferred learning styles (Chowdhury & Le Ha, 2008). CLT is an approach in language teaching that evolved in 1970s and is still one of the prevailing language teaching approaches today. One of the most characteristic features of CLT is that it aims at making communicative competence the goal of language teaching (Richards & Rodgers, 2001). Students in the Listening and Speaking courses are often put under great pressure for spontaneous language production and communication through role-play, small group work, and other CLT activities. As argued by Al-Arishi (1994), however, the specificity of role-playing in trying to replicate distinct situations in the real world "may work against the general communicative needs of the students, resulting in the students' seeing the tress, but missing the forest" (p. 340). In other words, the very specificity of CLT activities in listening and speaking classes may hinder language transferability to

real world situations and diminish the relevance of the activities to students' real needs and expectations. As argued by Wajnryb (1990: 14), "More and more we (TESOL teachers) are coming to realize that a methodology that violates the learners' preferred learning style will be of little value to them in the long run". Because of this sudden shift in style from traditional, clearly defined teacher-centred activities (students are used to) to CLT activities, students may become suddenly detached from their pre-college learning habits or preferred learning styles. This may leave them with a great degree of uncertainty and doubt over what constitutes good or bad teaching and how to judge teaching and teachers in these 'unfamiliar' waters even if they enjoy the new activities.

As can be seen in Table 8.2, classes taking the Writing Skills, Reading Skills, and the Core Course (integrated skills) enjoyed moderate to high inter-rater reliability in all the 8 subscales and the two overall rating items, with the exception of the Learning/Academic Value scale for the Core Course (r=.66). However, classes taking the Listening and Speaking Skills course yielded low (less than .70) inter-rater reliability in 6 out of the 8 scales. These six scales are: Learning/ Academic Value (r=.54), Organisation/clarity (r=.61), Group Interaction (r=.65), Individual Rapport (r=.47), Breadth of Coverage (r=.44), and Assessment/Grading (r=.64).

From the coefficients above, it appears that the Individual Rapport and the Breadth of Coverage are two dimensions of teaching that are difficult for students in listening and speaking courses to evaluate. As pointed out earlier, students' understanding, or lack of it, of their role as college students and the rules and regulations governing their relation with the college and their teachers, along with the cultural properties of friendliness held

by the students, may affect how students view and evaluate a teacher's rapport with his/her students. As for the Breadth of Coverage scale, the source of poor inter-rater reliability of students' ratings could be attributed to the fact that this scale is more applicable to content-based subjects than skill-based subjects.

In addition, because the teaching/learning activities in Listening and Speaking classes in the GFP are more CLT-oriented compared to the Reading and Writing Skills classes, lecturer's role as a facilitator, rather than a conveyer of knowledge, becomes more prominent. Here, teachers try to be more experimenting and creative. However, depending on teachers' knowledge, expertise, and level of preparedness, teaching styles and the selection of teaching/learning activities may vary greatly from one lesson to another with varying degrees of success. Therefore, listening and speaking teachers' techniques may become more difficult to judge and more susceptible to controversy among students. Role-play, for instance, may be suitable in short courses for intermediate and advanced adults and adolescents, but may not be suitable for long-term language programs because such activities may become meaningless and lacking in the cognitive and intellectual dimension, which is a very important motivating force for mature learners (Brumfit, 1984). As argued by Al-Arishi (1994), "Students … are obsessed with the cognitive and intellectual dimensions of language learning, and an activity which propels them to stray too far from those dimensions and which involves them with too many ungivens becomes meaningless" (p.342).

In summary, the evidence for the inter-rater reliability of the SEEQ scales in Oman is good. Although the inter-rater reliability estimates were found to improve as students

progress through the GFP levels, SEEQ proved to be reasonably reliable even when used with the Elementary Level students. No major differences in inter-rater reliability were found between classes differing in course type. The only exception is the Listening and Speaking Skills groups where SEEQ's inter-rater reliability was low in six out of the eight scales. However, the low estimates in this course can in part be attributed to the nature of the course itself and the ongoing controversy over some types of teaching/learning activities usually associated with it.

The content of some items in SEEQ would probably need to be revised if the instrument were to be used in collecting students' ratings from the GFP students, such as the items in the Breadth of Coverage scale, which are more relevant to content-based subjects. Some items may also need to be added to the Assessment/Grading and Assignments/Readings scales to accommodate subject-specific assessment techniques. This will not only make these two scales more applicable to the GFP programs, but it will also improve the power and stability of the two factors underlying them. Usually, a factor with three items or less can be weak and unstable (Costello & Osborne, 2005).

Nevertheless, bearing in mind the fact that all students sampled in this study were freshmen, the findings on the reliability of the SEEQ in the Omani context should be viewed as evidence of college students' ability to produce reliable ratings of their lecturer's teaching performance. The findings should not be seen as a call to use the instrument in other subjects, academic levels or disciplines in Oman's higher education institutions without further investigation. Follow up studies and trials of the instrument, including higher level students and covering a wider range of subjects, would yield more

generalisable findings on the reliability of this instrument in other subjects and academic levels.

8.3 Research Question 7: To what extent do student, lecturer, and course background characteristics influence students' ratings?

As discussed in Chapter 4, it is often suspected that certain background variables unrelated to teaching effectiveness may influence student ratings. In the SET research, this is what Marsh (1984) referred to as the witch hunt for potential biases in student' evaluations. Results from numerous studies have shown that various background characteristics are correlated with student ratings. In most studies, these background characteristics are grouped under three categories: student characteristics, teacher characteristics, and course characteristics (e.g. Aleamoni & Thomas, 1980; Cashin, 1995; Cenra, 1993; Kwan, 2000; Martin, 1998; Marsh, 1982b; Marsh, 1983). The size, significance, and nature of the relationship, as well as how it can be interpreted, however, remain an area of disagreement and debate. Some researchers (e.g. Feldman, 1997; Marsh, 1982b; Marsh, 1987; Marsh, 2007; Marsh & Dunkin, 1997, Marsh & Roche, 1997) pointed to a number of methodological problems in the research on potential biases in SETs. Implying causation from correlation in studying potential biases, for instance, is methodologically flawed (Centra, 1993; Marsh, 2007). Centra (2003), Feldman (1997), and Marsh & Dunkin (1997) also argued that poorly articulated operational definitions of bias and neglecting the multivariate nature of students' ratings has fuelled a lot of "myths" about bias in SET. As Marsh (2007) puts it:

> Recognition of the *multidimensionality* of teaching and of SETs is fundamental to the evaluation of competing interpretations of SET relations with other variables. Although a construct validity approach is now widely accepted in evaluating

> various aspects of validity, its potential usefulness for the examination of bias issues has generally been ignored.
>
> (p.347)

Centra (2003), Centra and Gaubatz (2000), and Marsh (1987) seem to agree on one operational definition of bias in SET: "Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning" (Centra, 2003: 498). Marsh (2007), however, stresses that if a potential background characteristic actually does have a valid influence on teaching effectiveness as evident in various measures of effective teaching (e.g. SETs, student motivation, test scores), "then it may be possible that the influence reflects support for the validity of SETs … rather than a bias" (Marsh, 2007: 349). Marsh (1984, 2007) also argue that if a potential biasing factor has a substantial effect on specific SET dimensions to which it is most logically related (e.g. class size and individual rapport), but has little or no effect on other dimensions of SET, this influence may also be taken as evidence of the validity of SETs rather than a bias.

The position taken in the present study with regard to exploring for potential biases in SET is similar to that of Marsh (1984) and Theall & Franklin (2001). This position entails that student ratings could be better understood and appreciated if researchers did not focus entirely on the witch hunt for potential biases or on trying to refute the existence of a bias, but instead carefully studied the nature and meaning of specific relations between background characteristics and students' ratings. In other words, "The simplistic bias hypothesis is a straw man and its rejection does not mean that student

ratings are unbiased, but only that they are not biased according to this definition" (Marsh, 1984: 733).

The SEEQ ratings collected for the present study were examined for such relationships with various student, lecturer, and course characteristics. Mann-Whitney U Test and Kruskal-Wallis Test were used to test for differences in lecturer overall ratings among different groups of students, lecturers, and courses. In addition, Spearman's Rank Order Correlation (rho) was used to examine correlations between overall ratings and other student, lecturer, and course characteristics. Table 8.3 below presents a summary of the background characteristics which were tested for relationships with students' overall ratings.

Table 8.3: Background Variables Tested For Relationship with Students' Overall Ratings

Student Characteristics	Lecturer Characteristics	Course Characteristics
Gender	Gender	Course type
GFP Level	Ethnic background	Course difficulty/ workload
Prior interest in the subject	First language	
	Grading leniency	

The results of the analysis for each of the above listed background variables, along with discussions of the findings are presented in the following sections.

8.3.1 Relationship with Student Characteristics

Overall teacher ratings were tested for relationships with student's gender, GFP Level, and prior interest in the subject.

8.3.1.1 Student's Gender

As shown in Table 8.4, a Mann-Whitney U test revealed statistically significant differences between male students (n= 577, *Md*= 4.05, *Mean Rank*= 439.55) and female students (*n*= 345, *Mdn*= 4.19, *Mean Rank*= 498.20), *U*=86870, *z*= -3.237, *p*=.001, *r*=.11, in the overall ratings they gave to their teachers.

Table 8.4: Differences in Teacher Overall Ratings between Male and Female Students

Grouping variable	Sub-groups	N	Median	Mean Rank	U	Z	Sig.	*r*
Student's Gender *P*<.05	Male	577	4.05	439.55	86870.000	-3.237	.001	.11
	Female	345	4.19	498.20				

Looking at the medians and mean ranks, it can be seen that female students gave higher overall ratings to their teachers than male students. Using Cohen (1988) criteria for effect size (.1=small, .3=medium effect, and .5=large effect), however, the *r* value found here, .11, is considered a small effect size and, therefore, no further analysis was carried out on this variable.

Many of the studies that compared the SET ratings of male and female students found essentially no difference between the groups (e.g. Aleamoni & Thomas, 1980; Feldman, 1977; Fernandez & Mateo, 1997; Ludwig & Meacham, 1997). The small effect size for student's gender on SET found in this study is typical of the many studies that have been carried out over the years and found little or no effect for student's gender on the evaluation of teaching effectiveness (Marsh, 1984, 2007).

The finding that female students gave higher ratings to their teachers than did the male students is also consistent with the findings reported in some old studies (e.g. Feldman, 1977) and in some more recent ones (e.g. Walumbwa & Ojode, 2000). In two subsequent meta-analyses, Feldman reviewed a number of studies that investigated the effect of gender- of both the teachers and the students- on students' ratings. Some of the studies were conducted in simulated settings or laboratories (Feldman, 1992) while others were conducted in actual classrooms (Feldman, 1993). In the laboratory studies, he found either no differences or inconsistent differences between different gender students' evaluations of their teachers. In three laboratory studies, however, male students were found to give lower ratings to female teachers. In the classroom-based studies, the findings were slightly different. Students rating a same-gender teacher recorded the highest ratings. Ratings given by female students to male teachers or by male students to female teachers, however, were lower.

Clearly, the findings about the effect of student's gender on students' ratings reported in the research cited above are mixed and inconclusive. More in-depth investigations are needed to examine interactions between student's gender and other attributes, such as personality traits, attitudes, motivation, and values to determine the extent to which differences in students' ratings between genders are influenced by such attributes.

8.3.1.2 Student's GFP Level

There seems to be no research on the effect of freshmen's background characteristics on SET. No research to examine the effect of English proficiency level on students' ratings could also be cited. Obviously, the research findings about the effect of student's level

on SET discussed in chapter 4 are mixed and inconclusive. Moreover, the implications of these findings may seem unclear for English language teaching because the term *level* in TESOL programs has a different meaning and is usually associated with language proficiency rather than the educational level. Nevertheless, even in language foundation programs, moving from one *level* to another denotes progress over time and the accumulation of experiences and skills as students move from one phase of the foundation year to the next. For the GFP students in the present study, a level takes a minimum of one semester to complete. The majority of the students enrolled in the foundation year start at the Elementary level and take a minimum of three semesters to complete up to the Advanced level. Bearing in mind the time spent in the GFP and the intensive nature of its courses, the foundation year can be very transforming, both in language proficiency and educationally. That is why student's GFP level, both as a unit of time spent in the college and as an indicator of language proficiency, is being investigated for effects on SET in this study.

The effect of students' GFP Level on students' overall rating of their teachers was investigated using Kruskal-Wallis test. As shown in Table 8.5, the Kruskal-Wallis test revealed no statistically significant differences in overall ratings given to teachers across the three GFP Levels (Elementary Level: n= 185, *Mdn*= 3.94, *Mean Rank*= 422.01; Intermediate Level: n=362, *Mdn*=4.25, *Mean Rank* =476.09; Advanced Level: n= 375, *Mdn*= 4.19, *Mean Rank*= 466.90), χ^2 (2, n=922)= 5.313, p= .07.

Table 8.5: Differences in Lecturer Overall Ratings across Different GFP Levels

Grouping variable	Sub-groups	N	Median	Mean Rank	χ^2	df	Sig.
Student's GFP Level *P*<.05	Elementary	185	3.94	422.01	5.313	2	.07
	Intermediate	362	4.25	476.09			
	Advanced	375	4.19	466.90			

Although it appears from the median and mean rank that Intermediate Level students gave slightly higher overall ratings than Elementary and Advanced Level students, the difference was small and not statistically significant. As a result, no post-hoc tests were required between the different pairs of Levels.

The effect of students' academic level (i.e. freshman, sophomore, graduate, etc.) on students' evaluations has been the subject of many studies. In many old studies reviewed by Feldman (1977) and some more recent studies (e.g. Aleamoni & Hexner, 1980; Morgan & Davies, 2006), no relationship was found between students' class year and teacher ratings. In other studies, however, higher level courses like graduate courses received slightly higher ratings (Aleamoni, 1981; Braskamp & Ory, 1994; Feldman, 1978; Moritsch & Suter, 1988). In SEEQ research, it was also observed that higher level courses tended to receive somewhat higher ratings (e.g. Marsh & Overall, 1979, Marsh & Hocevar, 1991b). Langbein (1994) attributes high students' ratings in higher level courses to students' high motivation, maturity and discriminating ability.

There were still others who found that freshmen and third-year students enjoyed higher ratings compared to middle-level students (Cranton & Smith, 1986; Koh & Tan, 1997). Koh and Tan argue that first year subjects receive better SET ratings because of their

relative ease and introductory nature. Contrary to these two studies, Badri *et al.* (2006) found that less favourable SET scores are associated with first- and third-year courses.

Clearly, the findings discussed above are mixed. Probably student's level in the GFP does have an effect on SET, especially at the Elementary Level where students' English proficiency may pose a challenge. "Because of limitations of language or cultural inhibitions, these students may be unable or unwilling to communicate as freely as native speakers would in written or oral forms of faculty evaluation" (Pennington & Young, 1989: 629). For summative purposes in ESL faculty evaluation, Pennington & Young proposed the following guidelines for the use of SET:

> - The instruments and procedures should be constructed by evaluation specialists sensitive to the nature of the ESL context.
> - The instruments must provide opportunities for responses other than choices on rating scales.
> - Students need to be oriented to the content and purposes of the evaluation instruments and procedures.
>
> (p.630)

The last point is particularly important. While limitations in language can be eased by presenting the rating instrument in students' mother tongue, limitations in knowledge and experience in using rating forms is an area which requires a lot of attention. Lack of proper induction programs for students on the use and purpose of students' ratings in lower levels may translate into "bias" if neglected. If students are to take the role of evaluators of teaching, then they must be trained and prepared for the job. Informing students about the utility of students' ratings in their departments and training them on how to use rating forms and what to look for when evaluating a teacher may help minimise the effect of course level on SET.

8.3.1.3 Student's Prior Interest in the Course

Student's prior interest in the course was reported in a number of studies as the most strongly related background variable to SETs (Feldman, 1977, 1978; Howard & Maxwell, 1980; Marsh & Dunkin, 1997). The effect of prior interest on SEEQ scores was found to be greater than the effect of any of 15 other background characteristics studied by Marsh (1980, 1983).

The relationship between student's prior interest in the course and teacher overall rating was investigated using Spearman's rho correlation coefficient (Table 8.6). There was a medium, positive correlation between the two variables, r_s = .43, n= 922, p<.001, with high levels of prior interest in the course associated with higher overall teacher ratings.

Table 8.6: Spearman's (Rho) Correlations between Prior Interest in the Course and Overall Ratings

Spearman's rho		**Overall Lecturer Rating**
Prior Interest in the Course	Correlation Coefficient	.432
	Sig. (1-tailed)	.000
	N	922

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

According to Cohen's (1988: 79-81) guidelines (small correlation: r_s =.10 to .29; medium correlation: r_s= .30 to .49; large correlation: r_s= .50 to 1.0), the strength of correlation between student's prior interest in the course and the overall rating given to the teacher is medium. The coefficient of determination was calculated by squaring the

$r_s$value (R_s^2) and then multiplying by 100 to convert the score to a "percentage of variance" (Pallant, 2007: 132). The R_s^2 indicated 18.66 per cent variance in the ranks shared by the two variables. In other words, student's prior interest in the course helps to explain 18.66 per cent of the variance in students' overall ratings of their teachers. This is quite a small amount of variance explained.

It is argued that higher student interest in the subject results in a more favourable learning environment, which in turn facilitates effective teaching and leads to high SET scores (Marsh & Dunkin,1997). This is being the case, the influence of prior interest on ratings should not be labelled as bias. Care should be taken, however, that the influence is not inherent to the subject matter, as this may represent a source of "unfairness" when the results of students' ratings are used for making personnel decisions (Marsh, 1987).

8.3.2 Relationship with Teacher Characteristics

Four lecturer background characteristics were investigated for association with overall students' ratings of their lecturers, namely: lecturer's gender, ethnic background, mother tongue, and perceived grading leniency.

8.3.2.1 Effect of Teacher's Gender on Overall Ratings

As shown in Table 8.7, a Mann-Whitney U test revealed statistically significant differences between male lecturers (*n*= 409, *Mdn*= 4.05, *Mean Rank*= 405.61) and female lecturers (*n*= 513, *Mdn*= 4.14 ,*Mean Rank*= 506.06), *U*= 82048, *z*= -5.692, *p*=.000, *r*=.19, in the overall ratings they received from their students.

Table 8.7: Effect of Teacher's Gender on Overall Ratings

Grouping variable	Sub-groups	N	Median	Mean Rank	U	Z	Sig.	*r*
Lecturer's Gender *P*<.05	Male	409	4.05	405.61	82048.000	-5.692	.000	.19
	Female	513	4.14	506.06				

While it is clear from the median and mean rank in Table 8.7 that female teachers received higher overall ratings from their students than their male counterparts, the effect size, .19, found here is a small one to warrant further analysis on this variable.

The findings from past research on the differences in students' evaluations between male and female teachers are also mixed. While in some studies male faculty received higher ratings than their female colleagues did (e.g. Basow & Silberg, 1987; Kierstead, D'Agnostino, & Dill, 1988; Sidanius & Crane, 1989), in other studies, female teachers recorded higher evaluations than their male counterparts (e.g. Feldman, 1993; Tatro, 1995). In a study carried out at an all-female United Arab Emirates university, male teachers were evaluated slightly higher than female teachers (Morgan & Davies, 2006). In twenty-eight studies of global ratings of teachers, however, the correlation between gender and overall rating of the teacher was found to be only .02, with female teachers receiving slightly higher ratings (Feldman, 1993). "These findings, though modest, coincide with gender-stereotypic behaviours; women are expected to be more caring and sensitive than men. Whether the female teachers actually exhibited behaviours or whether students merely expected the behaviours because the teachers were female is not known" (Centra, 1993: 75).

It is evident from past research, however, that male and female teachers tend to approach teaching in the classroom differently and judge the effectiveness of their own teaching differently (Stratham, Richardson & Cook, 1991). Among the differences observed by Stratham, Cook & Richardson were: a) women tended to regard students as the locus of teaching/learning activities, while men focused more on themselves as teachers; b) women placed more importance on interactive teaching styles and students' participation than men; c) women were judged more likable when they promoted active group interaction; and d) men received better competence and likability ratings when they adhered to their stereotypical masculine behaviours in the classroom and used a "teacher as expert" style.

Because students have different expectations about how men and women should behave in the classroom, students may react differently to male and female teachers and rate the "likability" and "competence" for male and female faculty on somewhat different criteria (Anderson & Miller, 1997). To control for this, Stratham *et al.* (1991) recommend that "In constructing evaluation instruments that measure specific behaviors, items tapping both types of behaviors ought to be included to avoid favoring one or the other approach" (p.152).

8.3.2.2 Teacher's Ethnic Background and Overall Ratings

A number of studies investigated the effect of race match/mismatch between teachers and students on students' gain and on teachers' perceptions and evaluation of their students (e.g. Dee, 2005; Ehrenberg, Goldhaber, & Brewer, 1995). Very few studies,

however, could by cited that investigated the effect of race, ethnicity, or nationality on students' evaluations of their teachers (e.g. Al-Issa & Sulieman, 2007; Reid, 2010).

Al-Issa & Sulieman (2007) carried out a study at the American University in Sharjah, United Arab Emirates, which investigated the perceived effect of a number of background variables on SET, including lecturer and student nationality. Among many other findings, they concluded that students' ratings are potentially biased by the teacher and student's nationality. Students' evaluations of their teachers were found to be influenced by the fact that the teacher is an Arab or Non-Arab. Moreover, among the various nationality groups represented in the student population (e.g. Levant, African, Indian sub-continent), "The evaluations of Gulf students were found to be most likely to be influenced by biasing factors" (ibid: 312-313). Another mediating factor cited by the researchers here is the language of instruction in high school. Students who had been instructed in Arabic in high school were found to be more influenced by bias in SET than students who attended high schools in which English or an Asian language was the medium of instruction.

In the present study, the association between students' ratings of teaching effectiveness and lecturer's ethnic background was investigated. A Kruskal-Wallis test revealed a statistically significant difference in overall ratings across the six different ethnic groups of teachers represented in this study, χ^2 (5, n= 922) = 184.896, p= .000. As shown in Table 8.8, African teachers received the highest median rating (4.47). With a slightly lower median (4.33), Omani Arab teachers recorded the highest mean rank (663.21) compared to the other five ethnic groups. The teachers with the lowest median (*Mdn*=

3.74) and the lowest mean rank (341.23), nevertheless, were those from Asian backgrounds (Indian, Pakistani, Philippine, etc).

Table 8.8: Differences in Overall Ratings for Teachers from Different Ethnic Backgrounds

Sub-groups	N	Mdn	Mean Rank	χ^2	df	Sig.
Omani Arab	109	4.33	663.21	184.896	5	.000
Non-Omani Arab	85	4.30	654.35			
Asian (Indian, Pakistani, Filipino, etc)	359	3.74	341.23			
White European	44	4.23	545.50			
White North American	292	3.95	464.24			
African	33	4.47	470.70			

To investigate whether there are significant differences between different pairs of ethnic groups and to determine the effect size of such differences, a number of post-hoc tests were carried out. Fifteen Mann-Whitney U tests between all possible combinations of ethnic groups were run (Table 8.9). To control for Type 1 errors, a Bonferroni correction to the alpha value was applied (.05/15= .003). The results of these post-hoc tests are presented in Table 8.9 overleaf.

From Table 8.9, it can be seen that seven statistically significant differences in teacher overall ratings in seven different pairs of ethnic groups were revealed. From examining the z scores, it is evident that the difference in overall ratings between Omani Arab lecturers (*n*= 109, *Mdn*= 4.33, *Mean Rank*= 348.77) and Asian (Indian, Pakistani, Filipino, etc.) lecturers (*n*= 359, *Mdn*= 3.74, *Mean Rank*= 199.81) was the biggest, *U*=7110, *z*= -10.082, *p*=.000, *r*= .47, indicating that Omani Arab lecturers received significantly higher overall ratings than their Asian counterparts.

Table 8.9: Post-Hoc Tests On the Effect of Teacher's Ethnic Background on Overall Ratings

Sub-groups	N	Mdn	Mean Rank	U	Z	Sig.	*r*
Omani Arab	109	4.33	107.36	3558	-2.787	.005	.20
Non-Omani Arab	85	4.30	84.86				
Omani Arab	109	4.33	348.77	7110	-10.082	.000*	.47
Asian (Indian, Pakistani, Filipino)	359	3.74	199.81				
Omani Arab	109	4.33	91.94	770	-6.692	.000*	.54
White European	44	4.23	40.00				
Omani Arab	109	4.33	261.18	9354	-6.362	.000*	.32
White North American	292	3.95	178.53				
Omani Arab	109	4.33	73.96	1530	-1.311	.190	.11
African	33	4.47	63.36				
Non-Omani Arab	85	4.30	339.78	5289	-9.381	.000*	.45
Asian (Indian, Pakistani, Filipino)	359	3.74	194.73				
Non-Omani Arab	85	4.30	70.95	1364	-2.550	.011	.22
White European	44	4.23	53.50				
Non-Omani Arab	85	4.30	267.16	5766	-7.525	.000*	.39
White North American	292	3.95	166.25				
Non-Omani Arab	85	4.30	63.60	1054	-2.121	.034	.19
African	33	4.47	48.94				
Asian (Indian, Pakistani, Filipino)	359	3.74	189.68	3476	-6.072	.000*	.30
White European	44	4.23	302.50				
Asian (Indian, Pakistani, Filipino)	359	3.74	285.31	37808	-6.123	.000*	.24
White North American	292	3.95	376.02				
Asian (Indian, Pakistani, Filipino)	359	3.74	191.69	4198	-2.774	.006	.14
African	33	4.47	248.79				
White European	44	4.23	201.00	4994	-2.385	.017	.13
White North American	292	3.95	163.60				
White European	44	4.23	38.50	704	-.234	.815	.02
African	33	4.47	39.67				
White North American	292	3.95	165.83	3991	-1.619	.105	.08
African	33	4.47	137.94				

* Significant at $p < .003$

The second most significant difference in overall ratings also involves Asian lecturers, but this time paired with non-Omani Arab lecturers. In this pair, non-Omani Arab lecturers were given significantly higher ratings (*n*= 85, *Mdn*= 4.30, *Mean Rank*= 339.78) compared to their Asian colleagues (*n*= 359, *Mdn*= 3.74, *Mean Rank*= 194.73), *U*=5289, *z*= -9.381, *p*=.000, r=.45.

The third biggest difference in overall ratings was observed between non-Omani Arab lecturers and white North American lecturers. Non-Omani Arab lecturers were rated significantly higher (*n*= 85, *Mdn*= 4.30, *Mean Rank*= 267.16) than white North American lecturers (*n*=292, *Mdn*= 3.95, *Mean Rank*= 166.25), *U*= 5766, *z*= -7.525, *p*= .000, r= .39.

Two other significant differences in teacher overall ratings between Arab teachers and white teachers were seen between Omani Arab teachers on one side, and White European and North American teachers on the other. The overall teaching effectiveness of Omani Arab lecturers was rated higher (*n*=109, *Mdn*=4.33, *Mean Rank*= 91.94) than that of their white European counterparts (*n*=44. *Mdn*=4.23, *Mean Rank*=40.00), *U*= 770, *z*= -6.692, *p*= .000, with a large effect size of *r*= .54. The Omani Arab teachers also scored higher in students' evaluations (*n*=109, *Mdn*= 4.33, *Mean Rank*= 261.18) compared to their white North American colleagues (*n*=292, *Mdn*=3.95, *Mean Rank*= 178.53), *U*= 9354, *z*= -6.362, *p*=.000, r=.32.

Unlike their fellow Omani Arabs, Asian teachers enjoyed lower overall ratings in front of their white North American and European colleagues. The overall teaching quality of Asian lecturers (*n*=359, *Mdn*= 3.74, *Mean Rank*=285.31) was rated lower than that of their white North American counterparts (*n*=292, *Mdn*= 3.95, *Mean Rank*= 376.02), *U*=37808, *z*= -6.123, *p*=.000, *r*=.24. The ratings of overall teaching effectiveness of Asian teachers (*n*=359, *Mdn*= 3.74, *Mean Rank*=189.68) was also lower than the ratings of white European lecturers (*n*=44. *Mdn*=4.23, *Mean Rank*=302.50), *U*=3476, *z*= -6.072, p=.000, *r*=.30.

To sum up, Arab GFP lecturers in general enjoyed significantly higher overall ratings compared to their Asian, white European, and white North American colleagues. White teachers, in turn, were given higher ratings by their students compared to Asian teachers. No significant differences in overall teaching effectiveness were found between African teachers and any of the other groups, but their ratings compared favourably in Mean Rank with the Asians and white Europeans.

It is not clear whether Omani students favoured Arab lecturers because they were from the same cultural and ethnic background or because they spoke the same language. Bearing in mind the fact that most students enrolled in the GFP had attended high schools in which Arabic was the medium of instruction and in which Omani and Arab teachers constituted the vast majority of teaching staff, it is probably not surprising to see Omani and Non-Omani Arab teachers receive the highest ratings. After twelve years of schooling, students become accustomed to their teachers' teaching style and any sudden changes may not be tolerated easily. This probably explains the findings by Al-Issa & Sulieman (2007) about the effect of high school language of instruction on students' perceptions of effective teaching. More findings and discussions about the effect of lecturer's mother tongue on SET are presented in the following section.

8.3.2.3 Teacher's First Language and Overall Ratings

The third teacher background characteristic tested for association with students' ratings of teaching effectiveness is lecturer's first language (L1). A Kruskal-Wallis test revealed a statistically significant difference in overall ratings across the six different languages represented in this study, χ^2 (4, n= 922) = 172.672, p= .000. As shown in Table 8.10,

native speakers of Arabic or Western Asian languages recorded the highest median, 4.32, in overall ratings of teaching compared to the speakers of the other four groups of languages. Teachers who spoke an Asian language from the Indian sub-continent or South Asia as their first language, on the other hand, received the lowest overall ratings in teaching effectiveness, median 3.74. When examining the mean ranks, however, one can see from the table that L1 speakers of Arabic and English received the best overall ratings in teaching quality (mean ranks 659.33 and 464.93 respectively), while L1 speakers of a South-East Asian language received the lowest SET ratings (mean rank 247.33).

Table 8.10: Differences in Overall Ratings for Groups of Teachers Differing in First Language

Sub-groups	N	Mdn	Mean Rank	χ^2	df	Sig.
Arabic	194	4.32	659.33	172.672	4	.000
English	347	3.95	464.93			
An Asian language from the Indian sub-continent or South Asia	305	3.74	364.87			
A Western Asian Language	40	4.32	401.85			
A South-East Asian Language	36	4.05	247.33			

In light of the above described significant differences across the five L1 language groups, and to investigate whether there are significant differences in overall teaching performance -as measured by SET ratings- between different pairs of L1 speakers, a number of post-hoc Mann-Whitney U tests were carried out. Ten Mann-Whitney U tests between all possible combinations of teacher L1 groups were run (Table 8.11). Again, to control for Type 1 errors, a Bonferroni correction to the alpha value was applied (.05/10= .005). The results of these post-hoc tests are presented in Table 8.11 overleaf.

Table 8.11: Post-Hoc Tests on the Effect of Teacher's First Language on Overall Ratings

Sub-groups	N	Mdn	Mean Rank	U	Z	Sig.	*r*
Arabic	194	4.32	353.24	17704	-9.157	.000*	.39
English	347	3.95	225.02				
Arabic	194	4.32	338.59	12399	-10.955	.000*	.49
An Asian language from the Indian sub-continent or South Asia	305	3.74	193.65				
Arabic	194	4.32	126.50	2134	-4.509	.000*	.29
A Western Asian Language	40	4.32	73.85				
Arabic	194	4.32	133.50	666	-9.564	.000*	.63
A South-East Asian Language	36	4.05	18.50				
English	347	3.95	365.41	39414	-5.630	.000*	.22
An Asian language from the Indian sub-continent or South Asia	305	3.74	282.23				
English	347	3.95	197.00	5898	-1.558	.119	.08
A Western Asian Language	40	4.32	167.95				
English	347	3.95	199.49	3646	-4.118	.000*	.21
A South-East Asian Language	36	4.05	119.78				
An Asian language from the Indian sub-continent or South Asia	305	3.74	172.92	6076	-.041	.968	.00
A Western Asian Language	40	4.32	173.60				
An Asian language from the Indian sub-continent or South Asia	305	3.74	175.07	4250	-2.221	.026	.12
A South-East Asian Language	36	4.05	136.56				
A Western Asian Language	40	4.32	47.95	342	-4.065	.000*	.47
A South-East Asian Language	36	4.05	28.00				

* Significant at $p<.005$ level

It can be seen from Table 8.11 above that seven statistically significant differences in lecturer overall evaluations of teaching in seven different pairs of L1 groups were revealed. The difference in overall teacher ratings with the biggest z score was observed between the teachers whose first language was Arabic (*n*=194, *Mdn*= 4.32, *Mean Rank*=338.59) and the teachers who spoke an Asian language from the Indian sub-

continent or South Asia as their first language (*n*=305, *Mdn*=3.74, *Mean Rank*=193.65), *U*= 12399, *z*= -10.955, *p*=.000, *r*= .49. It is clear here that teachers who spoke Arabic as their first language enjoyed higher ratings than those who spoke an Asian language from India, Pakistan or South Asia.

Native speakers of Arabic also ranked significantly higher in students' ratings compared to the native speakers of a South-East Asian language (*U*=666, *z*= -9.564, *p*=.000, r= 63), English (*U*=17704, *z*=-9.157, *p*=.000, *r*= .39), and a Western Asian language (*U*=2134, *z*=-4.509, *p*=.000, *r*=.29). Bearing in mind the nature of the GFP program, which is largely a TESL/TEFL foundation program, it is interesting to see that native speakers of English received lower ratings compared to those awarded to native speakers of Arabic.

Native speakers of English, however, were given higher ratings by their students compared to the teachers whose L1 was either an Asian language from the Indian sub-continent or South Asia (*U*=39414, *z*=-5.630, *p*=.000, *r*=.22), or a South-East Asian language (*U*=3646, *z*=-4.118, *p*=.000, *r*=.21). Teachers who spoke the latter as their L1 also scored slightly lower in students' ratings compared to the native speakers of a Western Asian language (*U*=342, *z*=-4.065, *p*=.000, *r*=.47).

As hinted at earlier in the previous section, teachers who are native speakers of Arabic, particularly Omani teachers, seem to receive the highest ratings in the GFP, even compared to native speakers of English. This is in contrast with the findings reported by Finegan & Siegfried (2000) and Ogier (2003) -discussed in chapter 3- who found that

non-native speakers of English were rated lower than native speakers of English and that a teacher's expressiveness in English had the highest correlation with the overall rating of the teacher.

Whether GFP students' preference for Omani teachers in the present study reflects a distinction in these teachers' quality of teaching, an advantage of being bilingual, or a mere student adaptation issue, remains to be seen. Probably one factor that needs to be considered here is the English language teaching (ELT) training of the teachers. From the researcher's own experience, the teachers with the most training in ELT in the GFP are Omanis. Most of them hold a minimum qualification of a BEd in TESOL and an MA/MEd in TESOL or applied linguistics. While native speakers of English teaching in the GFP have the advantage of being native to the language of instruction, they usually lack the proper training in TESOL, which probably explains the lower ratings given to them compared to the Omani and Arab teachers.

8.3.2.4 Teacher's Perceived Grading Leniency and Students' Ratings

The evidence in research of a relationship between grading and students' ratings is consistent, indicating a modest correlation of about .20 (Theall & Franklin, 2001). This relationship- also referred to as "grading leniency hypothesis" or "grading leniency effect" (Marsh, 1987; Marsh & Dunkin, 1992; Marsh & Roche, 2000)- has been the subject of a long debate between researchers in the field (e.g. Abrami & d'Apollonia, 1998; Greenwald & Gillmore, 1997; Marsh & Roche, 1998). Greenwald & Gillmore (1997) argue that there is a strong relationship between giving higher grades and obtaining high students' ratings. Abrami & d'Apollonia (1998) and Marsh & Roche

(1998) debated this view and stressed that the presence of this relationship in itself does not invalidate the established connection between ratings and learning. Cohen (1981) considered this relationship an expected phenomenon, as it reflects students' satisfaction level with learning. Using the same reasoning, McKeachie (1979) also argued that this relationship between grades and ratings is evidence of ratings validity rather than a sign of SET bias.

The relationship between lecturer's grading leniency as perceived by students and his/her overall rating was investigated using Spearman's rho correlation coefficient (Table 8.12). There was a medium, positive correlation between the two variables, r_s = .45, n= 922, p<.001, with high levels of grading leniency associated with higher overall teacher ratings. This correlation power is greater than what is reported by many studies in the field.

Table 8.12: Spearman's (Rho) Correlation between Perceived Grading Leniency and Overall Ratings

Spearman's rho		**Overall Lecturer Rating**
Perceived Grading Leniency	Correlation Coefficient	.445(**)
	Sig. (1-tailed)	.000
	N	922

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

The coefficient of determination was calculated and was found to be 19.80. In other words, the R_s^2 indicated 19.80 per cent variance in the ranks shared by the two variables, which is a modest amount of variance explained.

One important weakness in studies examining the link between grading leniency and teacher ratings is that few studies have incorporated measures of student perceptions of the teacher's grading leniency (Marsh, 1987; Marsh & Roche, 2000). In light of this finding, the relationship between grading leniency and ratings in this study was tested by correlating responses to a measure asking students to rate the perceived grading leniency of their teacher –rather than the expected or actual grade- with the overall rating. This decision was also driven by other practical considerations. The final grade in a GFP Level is communicated as a single composite grade incorporating all continuous assessment marks, quizzes, mid-semester exam results, and *Level Exit Exam* results obtained in all the four *skills* and/or courses surveyed together. In other words, the final grade is assigned to the Level as a whole, not to single *skills/courses*, although these courses are tested separately for continuous assessment purposes. Because of all of the above, no statistical analysis was carried out on item 35 in SEEQ (Appendix 9).

Clearly, the issue of whether the perceived easiness in obtaining good grades actually reflect students' high academic attainment or just teacher grading leniency needs to be considered carefully before accepting or dismissing the relationship as a bias.

8.3.3 Relationships with Course Characteristics

Three course characteristics were examined for association with teacher's overall ratings. These are: course type, course difficulty, course workload.

8.3.3.1 Effects of Course Type on Students' Ratings

The effect of course type on SET was investigated to determine whether there are any associations between subject matter and students' ratings of their teachers. A Kruskal-Wallis test revealed a statistically significant difference in the value of overall ratings across the four course types sampled in this study, χ^2 (3, *n*= 922) = 122.358, *p*= .000. As shown in Table 8.13, the Writing Skills course recorded the highest median and mean rank (*n*= 211, *Mdn*= 4.30, *Mean Rank*= 541.76) in overall ratings. The course with the lowest median and mean rank in teacher's overall rating, however, was the Reading Skills course (*n*=210, *Mdn*= 3.71, *Mean Rank*=286.49).

Table 8.13: Differences in Overall Ratings across Different Courses

Sub-groups	N	Mdn	Mean Rank	χ^2	df	Sig.
Writing Skills	211	4.30	541.76	122.358	3	.000
Reading Skills	210	3.71	286.49			
Listening and Speaking Skills	257	4.19	514.60			
Core Course	244	4.09	486.80			

To test for significant differences in SET ratings between the four courses, six post-hoc Mann-Whitney U tests were conducted. The alpha value was adjusted to control for Type 1 error using Bonferroni correction (.05/6= .008). The results of these tests are summarised in Table 8.14 overleaf.

As can be seen from Table 8.14, three statistically significant differences in overall ratings in three pairs of courses were revealed by the post-hoc tests. In all the three pairs, the Reading Skills course teachers appeared to receive lower ratings than the ratings of

Table 8.14: Post-Hoc Tests on the Effect Of Course Type on Overall Ratings

Sub-groups	N	Mdn	Mean Rank	U	Z	Sig.	*r*
Writing skills	211	4.30	268.48	10027	-9.727	.000*	.47
Reading skills	210	3.71	153.25				
Writing skills	211	4.30	246.35	24614	-1.719	.086	.08
Listening and speaking skills	257	4.19	224.77				
Writing skills	211	4.30	238.94	23434	-1.652	.099	.08
Core course	244	4.09	218.54				
Reading skills	210	3.71	165.97	12699	-9.857	.000*	.46
Listening and speaking skills	257	4.19	289.59				
Reading skills	210	3.71	178.27	15281	-7.425	.000*	.35
Core course	244	4.09	269.87				
Listening and speaking skills	257	4.19	258.23	29495	-1.149	.251	.05
Core course	244	4.09	243.38				

*** Significant at p<.008**

the teachers of the other courses compared with. Compared to the Writing Skills lecturers (*n*= 211, *Mdn*=4.30, *Mean Rank*=268.48), for instance, the Reading Skills lecturers (*n*=210, *Mdn*=3.71, *Mean Rank*= 153.25) were given significantly lower ratings by their students, *U*=10027, *z*=-9.727, *p*=.000, *r*=.47. The Reading Skills teachers (*n*=210, *Mdn*= 3.71, *Mean Rank*= 165.97) were also rated lower than the Listening & Speaking Skills teachers (*n*=257, *Mdn*= 4.19, *Mean Rank*= 289.59), *U*= 12699, *z*= -9.857, *p*=.000, *r*=.46. The same direction of difference was observed when Reading classes were compared to the Core Course classes.

This does not necessarily mean, however, that teaching quality varies across different courses. "What it does show is that effective teaching and learning may be harder to achieve under certain sets of conditions" (Theall & Franklin, 2001: 50). For instance, students in large science courses rating full-time faculty were found to give more

accurate evaluations compared to students in medium-sized language courses evaluating teaching assistants (d'Apollonia & Abrami, 1997).

It is quite surprising to see that GFP students gave their highest ratings to the Writing Skills teachers. Writing is a skill that is usually difficult to master even in students' mother tongue. From the researcher's own experience in the filed also, Writing Skills is the course in which students usually obtain their lowest marks in the GFP due to the difficulty inherent in training students to write for general or academic purposes in a second language. It can be argued, however, that writing courses are very rewarding intellectually and give a vivid sense of achievement and a measurable indicator of progress to students unlike reading courses, for example.

Bearing in mind the many differences in writing conventions between Arabic and English, it can be said also that the role of a teacher-as-expert in such courses assume a more prominent position. In a context where the perception of the teacher-as-expert is culturally rooted, and where students "have learned that somebody who is more qualified, more educated, and more expert than they in matters of education should be responsible for decisions relating to education" (Meleis, 1982: 443), the 'knowledge', advice, and support provided by a writing instructor are very much appreciated and very much sought after.

It can be concluded also that, unlike the teacher's role in receptive skills classes, i.e. reading and listening, the performance of the teacher in productive skills classes, i.e. writing and speaking, can be more observable and more measurable for the purpose of

students' ratings. This probably explains why listening and speaking classes ranked second in overall ratings after writing. It is difficult to tell, however, whether Listening & Speaking teachers received the second highest ratings because of their role in the listening or the speaking element of the course, as the two skills are taught together in this case.

8.3.3.2 The Effect of Course Difficulty and Workload on Students' Ratings

Course difficulty and workload are frequently cited as potential sources of bias in SETs, with less difficult courses and courses with lighter workload usually receiving higher SET ratings (Marsh, 2007). Marsh, however, argues that course difficulty/ workload is positively-not negatively- correlated with students' ratings. Other studies (e.g. Centra, 2003; Marsh, 2001; and Marsh & Roche, 2000) showed non-linear relationship between course difficulty/workload and overall ratings of teachers, with an inflection point where SETs levelled off and then decreased as workload increased.

The relationship between course difficulty as perceived by students and lecturer overall rating was investigated using Spearman's rho correlation coefficient (Table 8.15). A medium, negative correlation, r_s = -.47, n= 922, p<.001, was found between the two variables. High levels of course difficulty were associated with lower overall teacher ratings. This is clearly the opposite of what Marsh found.

The coefficient of determination was calculated and was found to be 21.90. In other words, the R_s^2 indicated 21.90 per cent variance in the ranks shared by the two variables.

Table 8.15: Correlation between Course Difficulty and Teacher Overall Rating

Spearman's rho		**Teacher Overall Rating**
Course Difficulty	Correlation Coefficient	-.468(**)
	Sig. (1-tailed)	.000
	N	922

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

Course workload influence on lecturer overall rating was investigated using Spearman's rho correlation coefficient (Table 8.16). A small, positive correlation, r_s = .19, n= 922, p<.001, was found between the two variables. Higher levels of course workload were associated with higher overall teacher ratings. This appears to be in agreement with Marsh's findings cited earlier.

Table 8.16: The Effect of Course Workload on Students' Ratings

Spearman's rho		**Lecturer Overall Rating**
Course Workload	Correlation Coefficient	.195(**)
	Sig. (1-tailed)	.000
	N	922

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

The coefficient of determination was calculated and was found to be 3.80. In other words, the R_s^2 indicated 3.80 per cent variance in the ranks shared by the two variables, which is a very small variance.

The implication is that higher course workload does not necessarily mean lower ratings. To the contrary, "…teachers in order to be good teachers- as well as improving their

SETs, should increase good workload, but decrease bad workload" (Marsh, 2007: 352). The positive correlation between high workload and high ratings, especially when examined against the negative relationship found earlier between course difficulty and SETs, also shows that students are capable of making a distinction between workload and difficulty. In other words, students may object to the inherent difficulty in the subject matter in some courses, but will still appreciate the value of good workload.

8.4 Chapter Summary

In summary, the factor structure of the SEEQ found in Oman partially fits the factor structure of this instrument found by Marsh in USA and Australia and by other researchers in other western countries. It is evident from the factor analysis that Omani students could clearly identify some of the dimensions of effective teaching underlying SEEQ. There is a degree of overlapping, however, between other SEEQ dimensions and these may not be separable in Oman. The Breadth of Coverage scale was the least interpretable in the whole matrix.

Five factors could be clearly identified with the teaching dimensions found in the western SEEQ research despite the few cross-loadings. Two of the eight factors included in the analysis, namely: Learning/Academic Value and Individual Rapport loaded independently and did not overlap with any other factors or had any cross-loading items. Three other factors- Lecturer Enthusiasm, Group Interaction, and Assignments/Readings- did load higher under the targeted factors, but had a cross-loading item or overlapped with non-target cross-loadings from other factors. The Organisation/Clarity factor overlapped partially with both, the Lecturer Enthusiasm and the Group Interaction

factors. Cross-loadings were also observed in the Assessment/Grading scale with the Assignments/Readings items.

Unlike the findings from America and Australia and other western countries, where the instructor overall rating item and the course overall rating item usually loaded higher under the Enthusiasm factor and the Learning/Academic Value factor respectively, both of the Overall Rating items in this study loaded higher under the Lecturer Enthusiasm factor. This gives an indication of the tremendous role enthusiasm plays in teacher's overall rating in the Omani context.

The least interpretable factor resulting from the factor analysis in this investigation is the Breadth of Coverage. None of the items loaded on its target factor. Instead, they loaded and cross-loaded with three other seemingly unrelated factors: Group Interaction, Lecturer Enthusiasm, and Assignments/Readings.

Omani students were also found to link teacher's personal and teaching skills (i.e. Enthusiasm and Group Interaction factors) with his/her organisational and planning skills (i.e. Organisation/Clarity factor), resulting in multiple cross-loadings and overlaps in the Organisation/Clarity scale with the other two factors. The overlap between these dimensions of teaching found in this study is similar to the findings reached by a number of researchers in various developing countries

Using SEEQ, the inter-rater reliability of students' ratings in Oman is good. The median inter-rater reliability coefficient for the eight sub-scales and the two overall rating items

is *r*=.87 for an average class size of 20 students. Although this coefficient is lower than the *r*=.89 found in Australia by Marsh & Roche (1993) and the *r*= .90 found in North America by (Marsh, 1987), the findings from Oman generally support the evidence about the inter-rater reliability of students' ratings. Although the inter-rater reliability estimates were found to improve with GFP levels, the reliability coefficients were reasonably good even at the Elementary Level. With the exception of the Listening and Speaking Skills classes where inter-rater reliability was relatively low in most of the scales, no significant differences in inter-rater reliability were found between classes differing in course type.

Teachers' overall ratings were tested for differences against three categories of background variables: student character tics (gender, GFP level, and prior interest in the subject), lecturer characteristics (gender, ethnic background, first language, and grading leniency), and course characteristics (course type, and course difficulty/workload).

Starting with student characteristics, it was found that female students gave higher overall ratings to their teachers than male students. No statistically significant differences in overall ratings were found between teachers across the three GFP Levels. There was a medium, positive correlation between student's prior interest in the course and teacher's overall rating.

Some teacher characteristics were also found to affect overall ratings. With a small effect size, it was found that female teachers received higher overall ratings from their students than their male colleagues. Also, Arab GFP teachers in general, and Omani

Arab teachers in particular, enjoyed significantly higher overall ratings compared to their Asian, white European, and white North American colleagues. White European and American teachers, however, were given higher ratings by their students compared to Asian teachers. No significant differences in overall teaching effectiveness were observed between African teachers and any of the other ethnic groups, but their ratings compared favourably with the Asians and white Europeans.

Teachers who spoke Arabic as their first language enjoyed higher overall ratings compared to those who spoke an Asian language from India, Pakistan or South Asia. Native speakers of Arabic also received significantly higher overall ratings compared to the native speakers of a South-East Asian language, English, or the natives of a Western Asian language. Native speakers of English, however, were awarded higher overall ratings by their students compared to the teachers whose mother tongue was either an Asian language from the Indian sub-continent or South Asia, or a South-East Asian language. Speakers of a South-East Asian language also scored slightly lower compared to the native speakers of a Western Asian language. Furthermore, there was a medium, positive correlation between high levels of grading leniency and higher overall teacher ratings.

As for the effect of course type on students' ratings, Reading Skills course teachers received the lowest ratings among the courses sampled. Writing Skills teachers, on the other hand, received the highest overall ratings. A medium, negative correlation was found between course difficulty and overall teacher ratings. Course workload, however,

was positively correlated with overall ratings. Higher levels of course workload seem to have resulted in higher overall teacher ratings.

CHAPTER NINE

SUMMARY, CONCLUSIONS, AND IMPLICATIONS FOR POLICY AND PRACTICE

9.0 Introduction

This final chapter presents a summary of the main findings of this study. The chapter begins by giving a brief overview of the problem under investigation. The chapter then presents a summary of the key findings presented and discussed in chapters six, seven, and eight. Based on these findings, conclusions are drawn and implications on the quality assurance policy and practice concerning the evaluation of teaching in general, and students' evaluations of teaching in particular, in higher education in Oman are identified. As this study is probably the first that explores the subject in hand in Oman, it should be recognised that these conclusions are preliminary and that more research will be required in the future to further explore some of the themes emerging from this study. Suggestions for future research are also given at the end of the chapter.

9.1 An Overview of the Study

Despite the widespread use of students' evaluations of teaching in higher education around the world and the huge body of research supporting their reliability, validity, and utility, especially for teaching improvement purposes, very few higher education institutions in Oman currently use systematic students' ratings in evaluating teaching. Students' "different" conceptions of quality teaching, "immaturity" and "inability" to give reliable and valid judgments about college teaching are some of the reasons often

given by teachers and administrators alike for not using students' ratings in the evaluation of teaching. None of these claims concerning the mismatch between the teachers and their students in their understanding of good teaching, or the lack of reliability and validity of students' ratings of teaching seem to have been substantiated by research evidence in Oman.

Probably due to this lack of trust in students' ratings, and more importantly, due to the lack of Oman-based research on SET, the practice remains largely unsystematic and in a state of almost total disconnect with the global research in the field. In the very few higher education institutions where students' ratings are collected, locally developed rating forms are usually administered without due consideration to the teaching constructs underlying them and without investigating the degree of match/mismatch between students' and teachers' perceptions of the characteristics of effective teaching underlying these instruments. In a context like Oman's higher education, where the students' population is largely homogenous, sharing the same cultural values, language, ethnic backgrounds, educational upbringing, and probably the same preferred learning styles, while their multi-national faculty is widely diverse in all of these aspects, the need for investigating the extent to which these two key stakeholders share the same meaning of effective teaching becomes even greater. This is a gap in Oman-based educational research this study attempts to bridge.

Furthermore, as the number of higher education institutions in the Arabian Gulf being modelled after or affiliated with western universities, most notably American, has increased sharply in the past few years, a growing concern among some researchers in

the region is that some of the western rating scales used in these institutions, are being adopted without due consideration to the new context. Part of the argument in the region has been that the cultural and educational upbringing of the students in the Gulf may affect students' perceptions of what constitutes effective teaching. This argument is also backed by some findings from other parts of the developing world where local students were found to mix certain dimensions of teaching underlying some American SET instruments. However, no research seems to have been carried out to investigate the transferability of these rating scales and their underlying constructs to the Gulf. More importantly, there seems to be no empirical evidence to establish or contest Arab students' ability to identify the multi-dimensions of teaching underlying rating scales. Therefore, part of this study comes in response to the lack of Gulf-based research in this area of SET.

All of the above is set against a background of educational change in Oman. The country has recently introduced a new quality assurance system in higher education with the student and the learning outcomes being the focus of the attention. In light of the above discussion about the lack of student involvement in the evaluation of teaching, it is argued that both HEIs and the new quality assurance policies need to recognise students' role in implementing and institutionalising the new quality vision. This role entails that students are involved in the evaluation of the teaching they receive in a systematic and effective manner. Without this involvement, various targets of quality assurance, especially those pertaining to teaching standards and professional development, may never by realised.

To this end, the present two phase study was carried out to provide evidence to policy makers, educators, and the many stakeholders of higher education on the potential of SET in Oman's higher education by investigating various aspects of students' ratings that currently seem to instigate some doubts and distrust in students' evaluations of teaching. Using the findings from the literature review and a qualitative exploratory study, a quantitative survey to identify teachers' and students' perceptions of the characteristics of effective teaching and SET was developed and administered to 248 teachers and 46 classes (968 students) in the GFP program in six colleges of technology in Oman. Two weeks later, SEEQ, the widely used American standardised SET instrument, was administered to the same 46 classes (922 students completed SEEQ) and each class was asked to rate one of their teachers chosen by the researcher. Data from the two instruments was then statistically analysed to:

- Identify the extent to which teachers' and students' perceptions of effective college teaching and students' evaluations of teaching matched or mismatched.
- Assess the association between students' and teachers' perceptions of effective teaching and various background variables.
- Identify the dimensions of teaching underlying students' evaluations of teaching in Oman and establish the extent to which these dimensions are similar to or different from those found in western countries.
- Assess the reliability of students' ratings in Oman and the effect of a number of course, teacher, and student background characteristics on these ratings.

It is believed that investigations like the one in hand are important not only in demystifying teachers' and students' perceptions of effective teaching, and the variables

affecting SETs and their potential in Oman's higher education, but also in helping college teachers in general, and foreign teachers working in Oman in particular, to better understand their students' expectations, preferences, and priorities in the college classroom. For those institutions which are currently collecting students' ratings, and for those now considering the use of students' ratings data in making decisions about the quality of teaching in their departments as part of the newly introduced national system for quality assurance in higher education, it is hoped that this well-timed study will offer them some insight in the subject.

9.2 Summary of Findings

Several important findings emerged from the data analysed in chapters six, seven, and eight. These findings are summarised below according to the themes derived from the research questions and research objectives and upon which the three chapters were organised.

9.2.1 Teachers' and Students' Perceptions of Effective Teaching

- A moderate but statistically significant overall correlation of *rho*=.56, $p<.001$ was found between teachers' and students' differential ranking of the importance of 38 characteristics of effective teaching. This correlation is slightly lower than what was found in some pioneering studies in the field.
- Teachers and students differed significantly in their rankings of the importance of 21 characteristics of effective teaching, but were closely matched on the other 17 characteristics.

- Five aspects of effective teaching were significantly more important for the GFP students: friendliness, pace of the lecture, building on previous knowledge, having native-like intonation and stress, and being a native speaker of the target language.
- Teachers ranked 16 characteristics significantly more important than did the students. These can be grouped under four main categories: challenging students, using student-centred approaches and maximising group interaction and students' input; teacher's academic qualifications, enthusiasm for the subject and the flexibility and diversity of his/her teaching styles; being supportive and available to students for advice, help, and feedback; and giving valuable and relevant homework and graded materials.
- Teachers' and students' rank-ordering of the significance of 17 other characteristics of effective teaching were closely matched. These can be grouped into 3 broad categories as follows: 1) preparedness for the job (mastery of the subject matter, keeping abreast of the latest developments in the field, sufficiency of experience and teacher training, and dedication to teaching); 2) preparation and presentation skills (good preparation of materials and use of teaching aids, lively and energetic presentation styles, expressiveness, and the ability to stimulate students' interest in the subject); and 3) Respect and fairness (respect for students, sensitivity to the culture of the organisation and society at large, fair evaluation, and compliance with the course objectives).

9.2.2 Mediating Background Variables Affecting Students' and Teachers' Perceptions of Effective Teaching

- Nine criteria of effective teaching showed significant differences in perceived importance between male and female students. These were: expressiveness, pace of the lecture, preparation, stimulating students' interest in the subject, being dynamic and energetic, fair evaluation, compliance with the course objectives, subject mastery, and being a native speaker of the target language. With the exception of the last one, female students assigned higher rankings to these characteristics than their male counterparts.
- Significant differences were found between the different ethnic groups represented in the teacher population in their perceptions of the importance of three dimensions of effective teaching: showing dedication to teaching, keeping abreast of the latest developments in the filed, and being a native speaker of the target language. In general, Asian teachers were found to attach greater importance to dedication to teaching and keeping abreast of the latest developments in the field compared to their European and North American colleagues. North American and European teachers, however, gave more weight to the nativeness of the lecturer to the target language (English) compared to their Arab and Asian counterparts.
- Significant differences were also observed between the perceptions of the teachers based on their first language in two aspects of effective teaching: having relevant academic qualifications in teaching English as a second/foreign language, and being a native speaker of the target language. L1 speakers of Arabic and Malayalam ranked the importance of academic qualifications

significantly higher than the L1 speakers of English or Urdu did. Native speakers of English, however, assigned significantly greater weight to being a native speaker of the target language compared to the teachers who spoke Malayalam, Tamil, or other South Asian and Southwest Asian languages.

9.2.3 Teachers' and Students' Perceptions of SET

- The teachers assigned significantly more weight to the effect of course workload and teacher's ethnicity on students' ratings.
- The students, however, placed significantly more prominence on the effect of lecturer's personal attributes and lecturer's mother tongue on student's ratings.
- Two of the factors hypothesised to bias students' ratings, lecturer's personal attributes and lecturer's communication skills, appeared to have the strongest effect on students' ratings from the teachers' and students' perspective.
- Teachers and students differed significantly on using SET data in making personnel decisions and in providing diagnostic feedback to teachers for improvement purposes, with students showing far stronger support for these two uses of students' ratings compared to their teachers.
- Teachers expressed significantly stronger support than the students did for limiting the use of SET results to teaching improvement only.
- Students expressed significantly higher confidence in their ability to evaluate most aspects of their lecturer's teaching performance compared to the teachers.
- Students also showed far more enthusiasm than their teachers did for working together with the teachers in developing rating forms and for learning about the characteristics of effective teaching.

- However, students were less enthusiastic than the teachers for receiving formal training on using rating instruments.

9.2.4 Students' Ability to Identify the Teaching Dimensions Underlying SEEQ

- The factor structure of SEEQ found in Oman partially fits the factor structure of the instrument found in America, Australia, and other western countries.
- Omani students could clearly identify five dimensions of SEEQ, namely: Learning/Academic Value, Individual Rapport, Lecturer Enthusiasm, Group Interaction, and Assignments/Readings, despite few cross-loadings and overlaps.
- Consistent with the findings of some SEEQ applicability studies conducted in the developing world, Omani students were found to link their teacher's personal and teaching skills (SEEQ's Enthusiasm and Group Interaction factors) with his/her organisation and planning skills (SEEQ's Organisation/Clarity factor), resulting in multiple cross-loadings and overlaps in the Organisation/Clarity scale items with the other two scales.
- The Breadth of Coverage factor was the least interpretable in the factor matrix resulting from the data collected from the GFP students in Oman. The items in this scale loaded and cross-loaded with three other seemingly unrelated factors: Group Interaction, Lecturer Enthusiasm, and Assignments/Readings.

9.2.5 Reliability of Students' Ratings in Oman

- The median inter-rater reliability coefficient for the eight SEEQ scales and the two overall rating items found in Oman is r=.87. It is slightly lower than the r=.89 found in Australia and the r=.90 found in North America.

- The inter-rater reliability of GFP students' ratings was found to improve with the GFP level, although the reliability coefficients were reasonably good even at the Elementary Level.
- With the exception of the Listening and Speaking Skills classes where the inter-rater reliability was relatively low in most of the scales, no significant differences in inter-rater reliability were found between classes differing in course type.

9.2.6 The Effect of Student, Teacher, and Course Background Characteristics on Teachers' Overall Ratings

- Female students gave higher overall ratings to their teachers than male students.
- A medium, positive correlation was found between student's prior interest in the course and teacher's overall rating.
- With a small effect size, it was found that female teachers received higher overall ratings from their students compared to their male colleagues.
- Arab teachers in general, and Omani Arab teachers in particular, received significantly higher overall ratings compared to their Asian, European, and North American colleagues.
- European and American teachers, however, were given higher overall ratings than the Asian teachers.
- Teachers who spoke Arabic as their first language enjoyed significantly higher overall ratings compared to the teachers whose first language was an Asian language or English.
- Native speaker of English, however, were given higher overall ratings by their students compared to the L1 speakers of Asian languages.

- A medium, positive correlation was found between high levels of grading leniency and higher overall teacher ratings.
- Reading Skills course teachers received the lowest overall ratings among the courses sampled, while Writing Skills teachers received the highest overall ratings.
- A medium, negative correlation was found between course difficulty and overall teacher ratings.
- A small, positive correlation was found between course workload and teacher overall rating.

9.3 Conclusions

Based on the findings of this study summarised above and discussed in more detail and in relation to the research literature in the filed in Chapters 6, 7, and 8, the following conclusions can be drawn. For ease of presentation and to aid reference to the relevant findings in the study, these are presented under the same themes under which the findings of the study were summarised in the previous section.

9.3.1 Teachers' and Students' Perceptions of Effective Teaching

While teachers and students in the GFP in the colleges of technology in Oman are closely matched in their perceptions of the importance of certain aspects of effective teaching, they differ significantly in their perceptions of other dimensions of quality instruction. Probably the area with the most significant mismatch between the two groups, and with the most serious implications, is their conception of challenging students, using student-centred approaches, and maximising students' participation and

input in classroom activities. While teachers seem to view this as a priority, students appear to be far less enthusiastic for this sort of approach in teaching/learning.

The implication is that teachers who are enthusiastic for student-centred teaching methods and who try to get the most out of their students and challenge them, whether in response to departmental mandates, as a personal preference, or in accordance with what is viewed as 'trendy' or contemporary in the literature on teaching methodology, may end up being punished by their students in the end of term ratings. Feeling unjustifiably punished, the teacher may be become demoralised and distrustful of the students and their ratings of his/her teaching. This in turn may add to the tension inside the classroom and send the relationship between the two key parties into a dysfunctional cycle of misunderstanding and punishment.

Therefore, the teaching quality standards embedded in quality assurance manuals, whether at the national level or the departmental level, should recognise not only what is internationally acceptable in the quality assurance circles and professionally advisable in the literature on teaching methods, but also what 'works' in reality depending on the context and the circumstances prevailing. Assuming that students will suddenly develop a taste for learner-centred approaches in the college classroom after 12 years of spoon-feeding is illogical. Even if students do not get to fill in rating forms for the courses they take, teachers may still be victimised by their heads of departments for conducting overly teacher-centred lessons.

When asked for their perceptions of the effective TESOL teacher, students were also found to show a preference for TESOL teachers who are native speakers of English or speakers with native-like intonation and stress. Although this preference was not precisely reflected in their actual overall ratings of their teachers as indicted in the findings summary earlier, this finding may pose a challenge to non-native speakers of English teaching in the GFP program who constitute the majority of the teaching staff. It is not clear whether students' preference for native speakers of English was because students faced difficulties understanding the speech of other L1 groups; because they enjoyed the classes of English native speakers more and benefited from them more; or because of pre-conceived notion that native speakers of English made better TESOL teachers. While the first and second reasons may be legitimate and valid, and based on observable teaching behaviours, the third reason clearly represents a bias against non-native speakers of English.

If students' perceptions of the effectiveness of native and non-native speakers of English are purely based on preconceived ideas, the implications on their attitudes towards their teachers whose first language is not English and the courses they teach could be very serious. Hard-working and committed non-native speaker TESOL teachers, with reasonable intonation and stress and good command of the English language, may be unfairly rated by their students. In a multi-national environment like the GFP program, this may not only damage the relationship between the students and their non-native English teachers and negatively affect students' attitudes, and possibly their attainment in their classes, but it may also damage the relationship between the native and non-native teachers themselves. With the program administrators being busy putting out fires

in a divided teacher community, the program's ability to successfully meet its objectives and achieve its mission becomes questionable.

Bearing in mind Oman's limited resources compared to the much richer Arab Gulf neighbours, Oman may find it increasingly difficult to attract high calibre native TESOL teachers. Therefore, the need for non-native English teachers from Asia and Africa and other parts of the world will continue in the future. Therefore, heads of departments and quality assurance officers should be prepared to tackle some students' irrational resentment to being taught by non-native speakers of English before it arises.

The finding that being supportive and available to students for advice, help, and feedback is more important for the teachers than for the students is very intriguing. One may be led to believe that the students do not realise the importance of the teacher's role in this regard or that students take their teachers' help and guidance in and outside class for granted. Looking at the bigger picture, however, this tendency on the students' part might be due to an interplay between students' culture and educational upbringing. As noted by Meleis (1982), Arab students have a strong need for affiliation and extensive social networking is an integral part of their everyday life. These social networks, Meleis adds, are considered a primary source of support, advice, and guidance in times of crisis. Meleis, however, stresses that "Neither the sharing of a problem nor advice are actively sought…Etiquette dictates that advice should be offered without a specific request" (ibid.: 441). In addition, because the current secondary education in Oman is largely teacher-centred (Issan & Gomaa, 2010), students usually enter higher education with the

expectation that college teachers will also be responsible for making all the 'right' decisions on their behalf.

Whether the rationale which makes students assign lower priority to seeking advice and support from their teachers is grounded in their social nature or educational nurture, the repercussions for this tendency can be strong. Unless the teacher is fully aware of the cultural and educational circumstances contributing to this stance, students' behaviour may be misinterpreted as lack of interest in the subject, or indifference to the teacher's efforts. Either way, it is a misunderstanding which may create a wide gap between the teacher and his/her students and deprive the students from valuable learning opportunities if left undetected.

One way to capitalise on Arab students' strong need for affiliation is to divide students into groups with team leaders selected from the most able and adaptable in the class. Such groups and team leaders can provide a good source of reassurance and support, and can be used as channels and networks through which teachers can offer their help. This technique is particularly suited for female students who, again for cultural reasons, usually refrain from meeting their teachers alone outside the classroom and insist on taking a friend or more with them to such meetings.

9.3.2 Mediating Background Variables Affecting Students' and Teachers' Perceptions of Effective Teaching

The finding that female students seem to attach significantly greater importance to seemingly unrelated dimensions of effective teaching, ranging from presentation and

teaching skills to personal skills, compared to their male classmates, is very interesting. Co-education in public schools in Oman exits in the first four primary grades only. After grade four, male and female students are segregated in separate schools, which are fully staffed by teachers of the same sex only, until they meet again in higher education institutions upon leaving grade 12.

In the absence of Oman-based research evidence, it is difficult to tell whether the observed differences between male and female students in the present investigation can be attributed to the difference in conditions under which male and female students were taught during the period of segregation in grades 5-12. Teaching styles and conditions seem to be the key variables here, as the educational system in Oman is highly centralised and other aspects like the curriculum, assessment, educational objectives, provisions and resources, and recruitment and staffing are all centrally managed by the Ministry of Education.

Because most higher education institutions in Oman are coeducational and staffed by both male and female teachers, unlike the segregated secondary schools, one would assume that the first year in college may pose a real adjustment challenge to both genders. In this year, the students are not only overwhelmed by the unfamiliarity of the new system they are entering and its different requirements, but also by the unfamiliarity of the classroom setting itself, where for the first time in their adolescence they are sharing a classroom with the opposite sex. This area, however, does not seem to attract much attention from either the teachers or the administrators in the colleges of technology. Usually, neither the teachers, nor the program administrators, are qualified

or prepared to deal with any adjustment problems that may arise as a result of this sudden reunion. Students are usually left to their own means in coping with any difficulties. This may reflect negatively on the performance of some vulnerable students from both genders.

More research on how first year students cope with co-education after segregation is needed. The findings from such research can offer college teachers, administrators, and students' affairs officers valuable input that may help them offer better and more relevant advice and counselling when the need arises. Such research can also give valuable insights on the uniqueness of each type of schools and the different experiences and skills that make students what they are when they join college.

Some demographic background variables seem to affect how some TESOL teachers perceive certain aspects of teaching effectiveness. Both teacher's ethnic background and mother tongue appear to have a role in determining his/her perceptions of the relative importance of qualifications and up-to-date knowledge in the field (TESOL) as opposed to being a native speaker of the target language. While Arab and Asian teachers and teachers with Asian first languages generally attached greater importance to TESOL qualifications and up-to-date knowledge in the field and far less importance to being a native speaker in English as a condition to teaching effectiveness, their European and North American colleagues and those teachers who spoke English as their first language appeared to take the position that being a native speaker of English is a key asset for an effective TESOL teacher.

If both parties fail to see the potential and strengths of each other and hold strongly to polarised views about the 'superiority' of either qualifications or nativeness to the target language, one of the implications for this could be that staff development programs may become extremely difficult to plan and implement, as their value and worth is likely to be contested or disregarded by one fraction or another. This could also result in continuous conflicts among the teachers and between the teachers and the program administrators.

9.3.3 Teachers' and Students' Perceptions of SET

Both teachers and students seem to think that some teacher's background characteristics have a strong effect on students' ratings. While teachers place more weight on teacher's ethnicity, for example, students seem to attribute stronger effect to lecturer's personal attributes, lecturer's mother tongue, and lecturer's communication skills on student's ratings. Again, ethnicity and mother tongue resurface as potential biasing factors to SET. It also appears that teachers are less enthusiastic than the students for using SET data in making personnel decisions, such as contract renewal or promotion. Compared to the students, they are also less confident in students' ability to evaluate college teaching. Students, however, seem willing to work with their teachers in developing rating scales and eager to learn more about the generic characterises of effective teaching.

It is interesting that both students and teachers seem to be aware of the mediating factors that can potentially bias students' ratings even in colleges where SETs have never been collected before. It is difficult to tell from the data available, however, whether teachers' position regarding the use of SET in making administrative decisions is mainly driven

by lack of trust in students' ability to evaluate their teaching or lack of trust in their administrators' interpretation and use of the ratings. Therefore, it is of paramount importance that teachers' concerns are listened to and examined carefully prior to deciding on the utility of students' ratings.

At the national level, quality assurance policies pertaining to teacher evaluation in general, and students' evaluations of teaching in particular, in higher education should be established on solid grounds of research evidence and not mere speculations. As such, these policies should take account of the huge body of research evidence supporting students' ratings and recognise students as an important stakeholder in the quality assurance process and as a viable source of data in evaluating teaching, side by side with other appropriate and well-researched means.

9.3.4 Students' Ability to Identify the Teaching Dimensions Underlying SEEQ

From the SEEQ factor structure found in Oman, it can be said that GFP students are capable of identifying most of the teaching dimensions underlying the instrument used in collecting their ratings. The SEEQ factor structure found in Oman largely, but not completely, replicates the factor structure of the instrument found in America, Australia, and other western countries. Despite few cross-loadings and overlaps, GFP students could clearly isolate five out of the eight dimensions of effective teaching underlying the American standardised SET instrument.

Some SEEQ factors, however, appear to be inseparable in Oman. Omani students seem to confuse certain aspects of their teacher's personal and teaching skills with his/her organisation and planning skills. Students in the GFP TESOL program also appear to be unable to isolate the Breadth of Coverage factor. Probably this particular scale is more appropriate for subject-based modules and not for skill-based courses like TESOL programs.

The findings here give a strong indication that students, even freshmen, seem to consider multi-dimensions of teaching in judging the performance of their teachers. While some extraneous factors like personal attributes may potentially affect students' ratings, students' approach to SET seems to be multi-dimensional and not overwhelmingly influenced by a single factor. While these findings do not answer the question whether these factors are the only important constructs of good teaching in the GFP, they do, however, give more reassurance that students' views on our teaching effectiveness are not based, at least not completely, on 'immature' or random observations, but on recognisable constructs of teaching.

9.3.5 Reliability of Students' Ratings in Oman

GFP students appear to be capable of giving reasonably reliable ratings of their teachers' performance in the classroom when using a well-designed SET instrument. Translated into Arabic for the first time and used with students who, for the most part, had never systematically rated a teacher before, SEEQ, a well researched standardised American SET instrument has proven to yield good inter-rater reliability coefficients, which are only slightly lower than those found in North America and Australia. The reliability of

students' ratings seem to improve as they progress through the Foundation year, with the ratings of the Intermediate and Advanced level students recording higher reliability coefficients than the ratings of the Elementary level students.

However, the low reliability coefficients in most of the SEEQ scales obtained from the listening and speaking skills courses raise the question whether satisfactory indices of inter-rater reliability of students' ratings may be difficult to obtain in courses where verbal communication skills are the foci of teaching/learning activities. It is not clear whether the poor level of agreement among students' ratings in these courses is attributable to an inherent difficulty in observing and rating the teacher's performance in such type of courses or because of the unfamiliarity of the nature of the course itself and the learning/teaching activities in it.

Unlike reading and writing skills, or integrated skills where listening and speaking are taught in conjunction with reading and writing, which are familiar modes of instruction and practice in secondary schools' English classes, listening and speaking skills taught in separate classes in college represent a complete novelty to students. Unlike the other classes in the program also, where the teacher still takes centre stage and a leading role as a transmitter of knowledge, listening and speaking classes, at least according to the announced objectives of the course, are characterised by increased student input and learner-centred activities. While students may enjoy these activities to the full, it may be more difficult for them to rate the less detectable contributions of their teachers using SEEQ scales as they are. This may require a bank of rating items developed for such courses which target specific teaching behaviours typical of the teacher-as-facilitator

role. Items from this bank can then be added to a core set of items in the SET rating questionnaire in use.

9.3.6 The Effect of Student, Teacher, and Course Background Characteristics on Teachers' Overall Ratings

It appears that students' ratings in the GFP are affected by various student, teacher, and course background characteristics. It must be noted, however, that these effects are based on correlational analyses and the analyses of differences between independent groups and do not necessarily reflect causal relationship between these background characteristics and overall teacher ratings. Therefore, careful interpretations should be drawn from these relationships and more research should be carried out which controls for various mediating variables before labelling any of these background variables as biases to SET.

From the data analysed, gender-both teachers' and students'-seems to be a factor against which some variations in ratings were observed. Female students' tendency to give higher ratings to their teachers and female teachers' lead in obtaining higher ratings from their classes compared to their male counterparts are very interesting findings. As far as female teachers' high ratings are concerned, probably a careful study of role expectations or gender-related preferences of teaching styles may help explain certain aspects of this relationship. For the female students, studying the culture of girls' schools in Oman or even the differences in upbringing between boys and girls dictated by the Arab culture may provide good answers to many questions and help explain some of the differences observed between male and female students in college classroom.

The positive correlation found between student's prior interest in the course and teacher's overall rating is consistent with the findings of many other studies in the field. In the context of the colleges of technology in Oman, however, this relationship between student's prior interest in the course and teacher overall ratings may pose very serious challenges. For many secondary school graduates in Oman, colleges of technology are the last choice after the other more popular government higher education institutions. Although this attitude has started to change recently, the fear is that those students who feel misplaced in the higher education system may fail to develop or maintain the required level of interest in their courses in CTs. This may reflect negatively in their ratings of their teachers. This calls into question the criteria used for admission and streaming of students into higher education programs in Oman and the potential effect these criteria may have on students' interest in their subjects at college.

The findings that Arabic speaking Omani teachers, Arab teachers and teachers who spoke Arabic as their first language received significantly higher ratings than did their colleagues from Asia, Europe, and North America, who spoke English or an Asian language as their first language, contradicts the findings of many researchers who concluded that in a TESL/TEFL context students tend to give higher ratings to native speakers of English. It is difficult to tell whether students' higher ratings given to Arab teachers here are the result of distinction in Arab teachers' teaching or the result of Arab students' strong need for affiliation discussed earlier. It may also be the case that Omani students are more accustomed to the teaching styles of Omani and Arab teachers, since most of the teachers in public schools are either Omanis or Arabs.

The other explanation which may pose a challenge to program administrators and non-Arab teachers alike is the possibility that Omani and Arab teachers use Arabic in the classroom in explaining difficult concepts and communicating complex ideas to their students or simply in translating vocabulary and instructions to lower level students. If the latter explanation is indeed the reason behind the high ratings awarded to Arab teachers, then there is a strong reason for concern that students' ratings may become largely determined by the teacher's first language, regardless of the teacher's performance in the classroom or students' attainment in the course.

The medium, positive correlation found between high levels of grading leniency and higher overall teacher ratings is an important finding. Trading marks for ratings is a serious threat to teaching standards and teacher's professionality and can diminish students' respect for their teachers and for the whole system. However, in practice, extreme care should be taken in interpreting the relation between marks and ratings. Effective teaching is usually expected to bring about improved learning and, consequently, improved performance in exams and better marks. Students may reward their effective teachers with good ratings in return. In this case, it can be argued that these ratings are earned legitimately and do not constitute a 'deal', but rather a valid indicator of good teaching. This exchange of marks and ratings may develop into a trading deal, however, when students' ratings of their teachers and grades spiral up in a dysfunctional cycle without underlying tangible improvement in either side. As with many aspects of education, proper training, effective leadership, quality assurance, and ethics of the profession should ensure that such practices remain the exception and not the rule.

What has been said above about grading leniency and overall ratings can also be said about the relation between course difficulty and ratings. The medium, negative correlation found between high course difficulty and high overall teacher ratings in this study may suggest that students tend to punish the teachers of difficult courses. It may well be the case, however, that difficult courses are difficult because of the teacher's lack of preparation and explanation skills, not because of reasons inherent in the subject itself or the level of work required.

There are two other findings from the present study which indirectly lend support to the argument given above. The first is that a small, positive correlation was found between course workload and teacher overall rating. This means that students may still reward an effective teacher with high ratings even if that teacher assigns a lot of hard work to students. The other finding showed that Writing Skills' teachers received the highest overall ratings, while Reading Skills course teachers received the lowest overall ratings among the courses sampled. From the researcher's teaching experience in the GFP, students usually view writing skills courses as the most difficult and most demanding of all courses in the program. These findings indicate that students can distinguish between difficulty and high workload and that while they seem to reject the first, they tend to appreciate the value of the latter. This is a further proof that simplistic explanations to the hypothesised biasing effect of some background variables on students' ratings do not hold and are better avoided.

9.4 Implications for Policy and Practice

Based on the findings of this study and the conclusions drawn from them, the following implications on policy and practice are identified:

1. The quality assurance system in higher education in Oman needs to recognise students as an important stakeholder in the educational process as a whole and in the evaluation of teaching and learning in particular. The decisions- national or institutional- on how students should be involved and what contributions they can make in the evaluation of teaching, for example, should be based on strong evidence and research-informed debates rather than speculations. Evidence from the present study shows that students can make reliable evaluations of their teachers. Findings also show that Omani students' are capable of identifying and separating various dimensions of teaching, which is evidence that their ratings are not overly dominated or determined by a single factor, personal or otherwise, but by a group of factors which share much in common with the factors identified by other college students around the world including some developed countries with a long history of SET, like USA and Australia.

2. Mandates and standards prescribing best practices in teaching and learning in Oman's higher education should be grounded on sound understanding of Omani students' educational upbringing in pre-college education. This is not to say that these mandates and standards ought to cement or condone the status quo in public schools' teaching approaches, but rather to recognise that some of the teaching strategies and methods students are used to may be in strong conflict

with the best practices in teaching emphasised in the country's quality assurance system for higher education. Teachers' strong enthusiasm for student-centred approaches coupled with students' deep-rooted preference for teacher-centred classes detected in this study, for instance, may have far reaching implications on college classrooms as discussed earlier. Students' switch from what is prevailing in high school to what is expected in college classrooms in quality manuals, however, should not be assumed to be automatic, spontaneous and trouble-free.

3. Addressing the gap between the skills emphasised by public schools and higher education expectations is a national dilemma, which requires long term strategies and immense efforts on the part of policy makers and policy implementers in both sectors, higher education and general education. The newly established quality assurance system in higher education should not underestimate the implications of this gap, nor should it assume that quality assurance in higher education can be fully and successfully implemented irrespective of the quality of pre-college education.

4. Until such national strategies bear fruit, identifying students' and teachers' perceptions and views of teaching and best practices in college classroom is an important pathway to addressing certain aspects of the problem highlighted above at the institutional level. The findings from such investigations, coupled with proper induction programs for teachers and students, can assist teaching evaluation and quality assurance programs achieve their goals and meet their benchmarks without triggering conflicts in college classrooms.

5. Given the position of the GFP program as a transitional stage between secondary education and higher education, GFP programs should strive to prepare their students for this shift in teaching style and educate them about the benefits of student-centred approaches in an overt, systematic and consistent manner. This may be integrated with the language course itself or as part of the study skills courses being offered.

6. Given the circumstance in secondary education hinted at above, GFPs should actively seek to gauge their students' views and perceptions about the quality of teaching in their programs, preferably through systematic students' ratings using well-constructed instruments. Such ratings will not only provide a good source of information in evaluating the quality of teaching and in planning their staff development programs, but may also reveal the effect of some mediating factors on students' perceptions of good teaching and help identify the degree to which their students' expectations match or mismatch with those of the teachers and the administrators. Where mismatches with serious implications are identified, steps should be taken at an early stage in the program to close the gap between students' expectations and needs by involving and engaging students in constructive dialogue and joint projects pertaining to the evaluation of teaching in the program. As evident in this study, students seem to be willing and interested in knowing more about the characteristics of effective teaching and working together with their teachers to reach shared definitions of good teaching.

7. Induction programs for teachers, especially those teachers who are not very familiar with the local culture, should not play down the potential effect of cultural differences on the relationship between teachers and students and their perceptions of the roles of each other. Probably assigning a mentor who is familiar with the local culture can help teachers who are new to the context cope with such culture-rooted misunderstandings.

8. GFP administrators and teachers should be ready to face the reality that students' evaluations of teaching, whether systematic or informal, may be influenced by a number of factors that may be considered irrelevant to effective teaching, such teacher's mother tongue, teacher's ethnicity, course difficulty, course type or other factors. Care should be taken, however, in interpreting these factors as biases to students' ratings, as there could be other underlying variables that may influence students' ratings which are not easy to detect from teacher overall ratings alone. These may stem from the nature of the course itself or even from teaching behaviours which are not covered by the rating instrument. In addition, some influences which may appear as biases may actually point to sources of validity in SET.

9. The native vs. non-native TESOL teacher debate seems to have arrived in Oman. While Oman needs the advantages both parties bring to the classroom in the GFP, the need for a more transparent and just recruitment system based on clear selection processes and criteria is even greater. Preferential treatment of candidates based on nationality or first language, regardless of qualifications and

abilities, could be very demoralising and may trigger a lot of resentment and conflicts among teachers. While it may not be always possible, program administrators should also try to consider their teachers' preferences and intrinsic strengths in allocating teaching duties. For instance, if a non-native TESOL teacher feels less confident teaching a speaking class, but shows great interest and potential in teaching a grammar class, then his/her wish should be considered. Staff development efforts should also highlight the strengths of both groups and encourage each group to learn from the potential and advantages of the other. At any cost, however, conflicts between native and non-native TESOL teachers should never be allowed to spill out into the classroom. The consequences could be disastrous, especially when students are known to show a preference for one type of TESOL teachers or another on irrational grounds like accent. This may not only affect students' ratings of teachers, but it may also discredit the whole practice and fuel even more suspicions about students' ability to give reliable and valid ratings.

10. From the literature review and the findings of this study, the use of students' ratings for personnel decisions seems to stir more controversy among teachers than using SET data for teaching improvement purposes. However, this controversy appears to be mainly driven by teachers' misunderstanding of how students reach their evaluations and what criteria they use. Therefore, when using SET in summative evaluation, it is extremely important that administrators engage in frank and professional dialogue with their teachers over this type of utility of students' ratings. Teachers should be told in clear terms how these

ratings will be used and for what type of decisions. A clear and confidential mechanism for collecting, analysing, interpreting, and reporting these ratings should also be put in place. Teachers and students should also be educated about the research findings on SET to increase their appreciation for the role of students' feedback in educational improvement.

11. In higher education institutions where SETs are introduced for the first time, the emphasis should be on using SET for teaching improvement purposes rather than personnel decisions. Both teachers and students need to see the effect of students' feedback on teaching improvement first before exploring with other administrative uses. Students should also be educated about the purpose of SET and the implications of their ratings on teachers and teaching. They should be trained how to use rating forms and told what each item means. In any case, using SET for teaching improvement also requires the development of systems and procedures for consultation over students' ratings, be it in groups or individually. Research has shown that individual or group consultation over students' ratings, for instance, improves teaching effectiveness.

12. For comparative purposes, the use of standardised rating instruments across departments and disciplines is advisable. However, such instruments should be flexible enough to accommodate rating items designed to probe certain teaching behaviours typical of specific courses. Different courses may develop and maintain a bank of items that can be added to a core set of items targeting generic dimensions of teaching. Whether standardised or designed to serve

specific courses, however, SET instruments should be robust and subjected to extensive testing to establish their psychometric properties before they are used to collect students' ratings.

9.5 Directions for Future Research

As hinted at earlier, because the present investigation is exploring an area that does not seem to have attracted much attention in Oman before, the study is bound to have limitations and its findings remain preliminary and in need for further verification and research in the future. One aspect which has not been investigated in this study is the effect of students' perceptions of effective teaching on their ratings of their teachers. Although this was proposed initially as one of the research questions in the present investigation, it became evident later that the scope of the study does not allow for such a major theme to be examined with the rest of the research questions in one small scale research project. It was also theorised that before looking at how students' perceptions of good teaching affect their ratings of their teachers' effectiveness, it would be more illuminating to examine students' conceptions of teaching effectiveness and the extent to which these conceptions match or mismatch with their teachers in the first place.

Another limitation of the study which has implications for future research in the filed in Oman is the generlisability of the sample. As indicated in the introductory chapter, the sample in the present investigation was selected exclusively from the GFP program of six colleges of technology in Oman to serve specific research objectives. While the research findings may still be relevant to other similar GFP programs in the rest of higher education institutions in Oman and the neighbouring Gulf states, it remains to be

seen whether such findings can also be replicated in other settings and with higher level students.

Another line of enquiry which can be picked up from the present study is the effect of matches/mismatches between students' and teachers' perceptions of effective teaching on students' attainment. A relevant area of investigation is also whether certain teaching approaches in Oman's colleges are more suited for Omani students and their educational background than others as judged by their results in standardised tests.

Future research can also make use of some of the data analysis procedures used in this study, such as factor analysis and inter-rater reliability analysis, to research or develop local rating instruments. Many of the rating forms used in Oman's colleges and universities are constructed without examining their psychometric properties and structural validity. Researching these instruments will provide relevant and more convincing evidence that can potentially be used to educate teachers and students about students' ratings.

Subsequent investigations may also examine the effect of educating students about effective college teaching and training them on using teacher rating forms on the factor validity and reliability of their ratings. Like any other evaluator or observer of teaching, students should be trained on how to make valid and reliable observations about teaching. Before resources for this training are allocated, however, evidence on the effect of such training on the quality of ratings should be collected. If training students as observers of teaching is proven to improve the quality of ratings in terms of validity

and reliability at least, then it may be argued that the GFP program-as an entrance program to higher education in Oman- is poised to play another important preparatory role that can potentially direct the debate from why involve students to how can we best prepare them for this involvement.

APPENDICES

APPENDIX 1

(Name of College)
English Language Center
(Staff Evaluation Form)

A: Lesson Evaluation Form

Academic year: ……. Semester: …… Date: ……
Lecturer: ……. Course: ……..…. Group: ……
No. of students present: …….. No. of students absent: …Period & time:……

	Evaluation Criteria	Rating				Comments
		Excellent	Good	Satisfactory	Requires Improvement	
I	TIME MANAGEMENT					
1	Punctuality					
2	Pace of lesson					
3	Use of class time					
II	LESSON DELIVERY					
4	Voice					
5	Nonverbal communication					
6	Confidence with students					
7	Teacher's enthusiasm for lesson/students' learning					
8	Varied use of techniques and exercises					
9	Evidence of lesson preparation					
10	Clarity of objectives					
11	Examples, demonstrations and illustrations meaningful and relevant					
12	Effective use of teaching aids					
13	Knowledge of subject area					
14	Students actively involved					
III	CONTINUED ASSESSMENT					
15	Lesson related to previous lesson, knowledge or interests					
16	Opportunities provided for language practice					
17	Appropriate reinforcement of learning given					
18	Effective error correction					
19	Previous lesson reviewed					
20	Observed individual differences					

IV	Classroom Management					
	Evaluation Criteria	Excellent	Good	Satisfactory	Requires Improvement	Comments
21	Instructions simple and clear					
22	Interruptions minimally disruptive					
23	Effective seating arrangement					
24	Students know what is expected of them and behave accordingly					
25	Attendance taken					
V	Student/Teacher Relations					
26	Rapport with students					
27	Sensitivity to learner needs					
28	Mutual respect					

Overall Rating:				
	28	28	28	28

B: Documentation	
Maintaining proper course information file	Score: /10

Other Comments (if any) :

Lecturer's signature: ……..……..

Observation Panel: Head of C&A Head of ELP's ELC Director

Signatures: …… …… ……

*A copy to be given to the lecturer after the feedback session.

C: Student Feedback Form

Lecturer's Name: ……. Group:……

Course: …… . Date: ……..

Please rate your teacher on the following points. 5 is highest, and 1 is lowest.

	Performance Indicators	5	4	3	2	1
1	The teacher starts classes on time. يبدأ محاضرتة في الوقت المحدد					
2	The teacher takes the attendance. يسجل الحضور والغياب					
3	The teacher has good control over the class. يستطيع التعامل والتحكم بالفصل بطريقة ممتازة					
4	The teacher gives clear explanations and examples. يعطي شرحاً واضحاً للدرس بالأضافة الى الأمثلة التوضيحيه					
5	The teacher encourages all students to learn. يشجع جميع الطلبة على التعلم والمشاركة					
6	The teacher is well organized. منظم في عملة أثناء المحاضرة					
7	The teacher's pronunciation is clear. وضوح اللغة واللفظ					
8	The teacher delivers the lesson at an appropriate pace. معتدل في تنفيذه للمحاضرة بطريقة تناسب جميع المستويات					
9	The teacher gives feedback to students. يعطي ملاحظاته وتوجيهاتة للطلبة بشكل مناسب					
10	The teacher gives time for students to practice their English. يمنح الطلبة وقتاً كافياً لممارسة لغتهم الأنجليزية					
11	The teacher listens and responds to students. يشجع ويستجيب لملاحظات ومشاكل الطلبة					
12	The teacher respects the students. يحترم الطلبة					
13	The teacher is fair. يتعامل مع الجميع بنفس المستوى					
14	The teacher is neatly dressed. أناقة المظهر					
15	The teacher ends the class on time. ينهي المحاضرة في الوقت المحدد					
16	The teacher's class is enjoyable. المحاضرة مشوقة وممتعة					

D: Lecturer's General Performance Report

#	Evaluation Criteria	Rating			
		Excellent	Good	Satisfactory	Requires Improvement
1	Attendance				
2	Punctuality				
3	Meeting deadlines				
4	Adherence to systems and regulations				
5	Participation in extracurricular activities				
6	Readiness for self improvement and keeping abreast with the latest developments in ELT field				
7	Relationship with direct senior staff				
8	Relationship with colleagues				
9	Observing society ethics				
10	Response to instructions				

Overall Rating:				
	10	10	10	10

Other Comments (if any):

……

……

ELC Director: Date: ……

E. Staff Appraisal Procedure

As part of the quality assurance system applied by the ELC and in conformity with the college bylaws, the ELC implements the following staff appraisal system:

1. The ELC conducts staff appraisal periodically, using the forms attached. These forms have been specifically designed for observing classes where English is taught as a second or foreign language. This practice reflects the differences between teaching a language and teaching other content-based subjects – the differences in methodology, materials, classroom management strategies, assignment of work for further language practice, etc.
2. The visits to the class will be scheduled by the ELC management and teachers may be visited without prior notice as and when required. However, teachers are usually informed in advance about the period during which their classes will be observed. A teacher may be visited a number of times as and when needed.
3. New staff are apprised / evaluated during the probationary period of three months from the date of reporting for work. Other members of staff are evaluated only if there is a serious concern about a particular teacher.
4. Two members of the observation panel will do the classroom observation.
5. Students are the beneficiaries of language instruction, so they also participate in staff appraisal by filling in a feedback form (attached), which includes statements in both English and Arabic.
6. The teacher observed is called for a feedback session. During the feedback session, the teacher can seek clarifications from the panellists regarding the criteria and the rating given for the teacher. He/She is also given a copy of the students' feedback.
7. As part of staff appraisal, the Course Information File maintained by the teacher will be assessed against the specifications outlined by the ELC.
8. At the end of this session, the teacher signs on the staff assessment form and a copy of it is given to the teacher concerned.
9. If a teacher refuses to sign the staff appraisal form, a note indicating this refusal will be added to the form by the evaluation panel.
10. The panellists make recommendations for action to be taken on areas that need further improvement and help is provided to the teachers concerned when required.
11. The original staff appraisal forms will be kept in the teacher's personal file and maintained at the office of the ELC Director. A summary of the evaluation data is sent to the Dean of the college along with an analysis of the data.
12. Based on the staff appraisal records, plans are drawn up for in-service programmes (i.e., workshops, seminars, presentations, etc) pertinent to the areas that need special focus. The staff are also consulted through memos with regard to topics for workshops, etc. External resource persons are sometimes invited to conduct staff development programmes.

APPENDIX 2

Data to Evaluate College Teaching (Adapted from Cashin, 1989)

	Sources of data in evaluating college teaching								
Areas of teaching	Self	Files	Students	Peers	Colleagues	Chair/Dean	Administrator	Consultant	Others
Subject matter mastery									
Content areas	a	bc		de		?			
Comprehensiveness	a	bc		de		?			
Currency	a	bc		de		?			
Objectivity	a	c		de		?			
Curriculum development									
Fit w/ other courses	a	c		de	de	de		?	
Course revisions	a	c		de		de			f
New courses	a	c		de		de			f
Course design									
Instructional goals	a	c	?	deg	?	deg		dg	
Content coverage	a	c	?	deg	?	?		dg	
Teaching methods	a	c	?	deg	?	deg		dg	
Assessment methods	a	c	?	de	?	deg		dg	
Delivery of instruction									
Methods	a		hij	deg	deg			dg	
Skills	a		hij	deg	deg			dg	
Aids	a	c	hij	deg	deg			dg	
Assessment of instruction									
Tests	a	c	hij	deg	deg	deg		dg	
Papers, projects	a	c	hij	deg	deg	deg		dg	
Practicums	a	c	hij	deg	deg	deg		dg	
Grading practices	a	c	hij	deg	deg	deg		dg	
Availability to students									
Office hours	a	k	hij	?		?			
Other	a		hij	?		?			
Informal contacts	a		hij	?		?			
Administrative requirements									
Book orders	a	l	hij			m	n		
Library service	a	l	hij			m	o		
Syllabi on file	a	l	hij			m	p		m
Comes to class	a		hij	?		?			
Grade reports	a	l	hij			m	q		

A **?** suggests the person(s) may be a source of data in some cases

a Self-report **b** Degrees, certificates, licences, etc. **c** Course materials on file **d** Review of course materials **e** Personal contact **f** Community advisory committee minutes, letters, etc.	**g** Classroom observation, video- or audiotapes **h** Student ratings **i** Student interviews **j** Students comments or letters **k** Posted office hours **l** Instructor's dated copies	**m** Department/Division/ College files **n** Book store manager **o** Librarians **p** Appropriate secretary **q** Register

APPENDIX 3

An Exploratory Study on the Evaluation of Teaching Effectiveness in the Foundation Programme in Colleges of Technology

ADMINISTRATOR QUESTIONNAIRE

Dear Colleague,

Thank you for agreeing to take part in this study. The purpose of this questionnaire is to investigate the teaching performance appraisal practices in the Foundation Programme in Colleges of Technology, and the role that programme administrators, teachers, and students play in the evaluation of teaching effectiveness. The findings of this survey will be used to identify the major themes and dimensions of the research instruments for the main study which will be conducted in the coming few months.

All data will be treated confidentially. Information obtained about you and the views you express in your answers will not be shared with your College; neither will your identity be disclosed in the research report.

Your participation in this study is voluntary. You may withdraw from the study at any time for any reason and without prejudice.

The questionnaire consists of three parts. **Part 1** asks for **background information** about you. Apart from your name, you are kindly requested to answer all the questions in this part. **Part 2** consists of **14 open-ended questions**. Space is provided for you to type the answer under each question. **Part 3** is a blank section for you to add any **additional comments** that you may have about the subject under investigation, or any other related issues you think are of importance to this study or that require further attention.

Completed questionnaires should be sent back to me as an e-mail attachment, to the following address:

nasser.alhinai@yahoo.com

Once again, thank you for taking part in this study.

Sincerely,

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

Part I:

1. Name *(optional)*:

2. College *(Please select one):*

1.	2.	3.	4.	5.	6.	7.
Muscat	Shinas	Musana	Ibra	Nizwa	Salalah	Ibri

3. Gender: 1. Male ( ) 2. Female ( )

4. Age group *(Please select one):*

1.	2.	3.	4.	5.	6.	7.	8.
Under 25	25-30	31-35	36-40	41-45	46-50	51-60	Over 60

5. Position/ Job title *(Please select one)*:

1.	2.	3.
Director of English Language Centre	Head of Section	Other: *(please specify)*

6. What is your first language? *(Please select one):*

1.	2.	3.	4.	5.	6.
Arabic	English	Urdu	Hindi	French	Other: (*please specify)*

7. Years of experience in English language teaching *(Please select a category):*

1.	2.	3.	4.	5.
0-5 years	6-10 years	11-15 years	16-20 years	More than 20 years

8. Ethnic background *(Please select a category)*:

1.	2.	3.	4.	5.	6.	5.
Omani Arab	Non-Omani Arab	South-western Asian	European	African	North American	Other: (*please specify)*

Part II:

Please answer the following questions:

1. In your opinion, is there a need for evaluating the teaching performance of faculty in higher education? Why? Why not?

2. What **factors** do you consider in **evaluating the overall performance** of your teachers (e.g. classroom teaching, experience in the field, committee work, personal attributes, etc.)? Please rank the factors you list, from the most frequently used to the least frequently used.

3. What **sources of information** do you use in **evaluating the teaching effectiveness** of your teachers (e.g. classroom visits, systematic student ratings, opinions from colleagues, etc.)? Please rank the sources you list, from the most frequently used to the least frequently used.

4. How much time do you devote to staff appraisal? Do you consider it to be time well-spent?

5. What role do students in your department play in evaluating the teaching effectiveness of their teachers?

6. Do you think that the Foundation Programme students are capable of judging the teaching performance of their teachers? Why? Why not?

7. In your opinion, what are the characteristics of effective college TESOL teachers?

8. What do you consider to be the prime purpose or goal of staff evaluations in your department?

9. What parts or practices of your current teaching performance appraisal system do you think have helped you most to achieve your teacher evaluation goals?

10. What parts or practices of your current teaching performance appraisal system do you think have hindered the achievement of your teacher evaluation goals?

11. How has your current staff appraisal system affected the quality of teaching/ learning in your department? Could you please give some examples?

12. What do you think are the contextual (legal, social, political, economic, and cultural) factors that influence the framing and implementation of staff appraisal in your department?

13. Which do you think is more appropriate for higher education institutions in Oman: staff appraisal models which emphasise accountability and quality control, or those which emphasise professional development and improvement? And why?

14. Do you consider teaching to be a labour, a profession, a craft, or an art? And why?

Part III: Any additional comments:

Thank you for taking the time to complete this questionnaire.

APPENDIX 4

An Exploratory Study on the Evaluation of Teaching Effectiveness in the Foundation Programme in Colleges of Technology

TEACHER QUESTIONNAIRE

Dear Colleague,

Thank you for agreeing to take part in this study. The purpose of this questionnaire is to investigate the teaching performance appraisal practices in the Foundation Programme in Colleges of Technology and the role that programme administrators, teachers, and students play in the evaluation of teaching effectiveness. The findings of this survey will be used to identify the major themes and dimensions of the research instruments for the main study which will be conducted in the coming few months.

All data will be treated confidentially. Information obtained about you and the views you express in your answers will not be shared with your College; neither will your identity be disclosed in the research report.

Your participation in this study is voluntary. You may withdraw from the study at any time for any reason and without prejudice.

The questionnaire consists of three parts. **Part 1** asks for **background information**. Apart from your name, you are kindly required to answer all the questions in this part. **Part 2** consists of **14 open-ended questions**. Space is provided for you to type the answer under each question. **Part 3** is a blank section for you to add any **additional comments** that you may have about the subject under investigation, or any other related issues you think are of importance to this study or that require further attention.

Completed questionnaires should be sent back to me as an e-mail attachment, to the following address:

nasser.alhinai@yahoo.com

Once again, thank you for taking part in this study.

Sincerely,

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

Part I:

1. Name *(optional)*:

2. College *(Please select one):*

1.	2.	3.	4.	5.	6.	7.
Muscat	Shinas	Musana	Ibra	Nizwa	Salalah	Ibri

3. Gender: 1. Male ( ) 2. Female ( )

4. Age group *(Please select one):*

1.	2.	3.	4.	5.	6.	7.	8.
Under 25	25-30	31-35	36-40	41-45	46-50	51-60	Over 60

5. Position/ Job title *(Please select one)*:

1.	Lecturer	
2.	Trainee Lecturer	

6. What is your first language? *(Please select one):*

1.	2.	3.	4.	5.	6.
Arabic	English	Urdu	Hindi	French	Other: (*please specify)*

7. Years of experience in English language teaching *(Please select a category):*

1.	2.	3.	4.	5.
0-5 years	6-10 years	11-15 years	16-20 years	More than 20 years

8. Ethnic background *(Please select a category)*:

1.	2.	3.	4.	5.	6.	7.
Omani Arab	Non-Omani Arab	South-western Asian	European	African	North American	Other: (*please specify)*

Part II:

Please answer the following questions:

1. In your opinion, is there a need for evaluating the teaching performance of faculty in higher education? Why? Why not?

2. Is there a system for staff appraisal in your department? If YES, what **factors** does your evaluator consider in **assessing your overall performance** (e.g. classroom teaching, experience in the field, committee work, personal attributes, etc.)? Please rank the factors you list, from the most frequently used to the least frequently used.

3. What **sources of information** does your evaluator use in **assessing your teaching effectiveness** (e.g. classroom visits, systematic student ratings, opinions from colleagues, etc.)? Please rank the sources you list, from the most frequently used to the least frequently used.

4. How far is staff appraisal in your department developmental and/ or evaluative?

5. What role do your students play in evaluating your teaching effectiveness?

6. Do you think that the Foundation Programme students are capable of judging the teaching performance of their teachers? Why? Why not?

7. In your opinion, what are the characteristics of effective college TESOL teachers?

8. What do you consider to be the prime purpose or goal of staff evaluations in your department?

9. What parts or practices of your current teaching performance appraisal system do you think have helped you to improve your teaching?

10. What parts or practices of your current teaching performance appraisal system do you think have negatively affected your teaching?

11. How has your current staff appraisal system affected the quality of student learning? Could you please give some examples?

12. What do you think are the contextual (legal, social, political, economic, and cultural) factors that influence the framing and implementation of staff appraisal in your department?

13. Which do you think is more appropriate for higher education institutions in Oman: staff appraisal models which emphasise accountability and quality control or those which emphasise professional development and improvement? And why?

14. Do you consider teaching to be a labour, a profession, a craft, or an art? And why?

Part III: Any additional comments:

Thank you for taking the time to complete this questionnaire.

APPENDIX 5

An Exploratory Study on the Evaluation of Teaching Effectiveness in the Foundation Programme in Colleges of Technology

دراسة أولية عن نظام تقييم أداء المحاضرين في البرنامج التأسيسي بالكليات التقنية

STUDENT QUESTIONNAIRE

استبيان الطالب

Dear Student,

Thank you for agreeing to take part in this study. The purpose of this questionnaire is to investigate the teaching performance appraisal practices in the Foundation Programme in Colleges of Technology and the role that programme administrators, teachers, and students play in the evaluation of teaching effectiveness. The findings of this survey will be used to identify the major themes and dimensions of the research instruments for the main study which will be conducted in the coming few months.

All data will be treated confidentially. Information obtained about you and the views you express in your answers will not be shared with your College; neither will your identity be disclosed in the research report.

Your participation in this study is voluntary. You may withdraw from the study at any time for any reason and without prejudice.

The questionnaire consists of three parts. **Part 1** asks for **background information**. You are kindly requested to answer all the questions in this part. **Part 2** consists of **11 open-ended questions**. Space is provided for you to type the answer under each question. **Part 3** is a blank section for you to add any **additional comments** that you may have about the subject under investigation, or any other related issues you think are of importance to this study or that require further attention. Completed questionnaires should be sent back to me as an e-mail attachment to the following address:
nasser.alhinai@yahoo.com

Once again, thank you for taking part in this study.

Sincerely,

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

أعزائي الطلبة و الطالبات:

بداية أتوجه بالشكر الجزيل لكم على مشاركتكم في الإجابة على هذا الاستبيان. كما سبق الإشارة يهدف هذا الاستبيان الأولي إلى دراسة نظام تقييم أداء المحاضرين في البرنامج التأسيسي بالكليات التقنية في السلطنة و إستطلاع أفكار و آراء طلاب و محاضري و إداريي البرنامج التأسيسي فيما يتعلق بكفاءة التدريس في هذا البرنامج و دور كل منهم في عملية تقييم أداء المحاضرين داخل قاعة الصف.

تستغرق الإجابة على هذا الاستبيان حوالي 20 دقيقة و يمكنكم طباعة إجاباتكم باللغة العربية أو الانجليزية في الأماكن المخصصة للإجابة تحت كل سؤال.

أود التأكيد هنا على أن جميع الإجابات و المعلومات التي ستدلون بها ستعامل بسرية تامة و لن يتم استخدامها إلا من قبل الباحث و لأغراض البحث فقط بهدف إستنباط محاور أسئلة لأدوات البحث النهائي الذي سيتم إجرائه في الأشهر القليلة القادمة. كما أؤكد على أن حقكم مكفول في الانسحاب من الدراسة في أي وقت تشاءون و بدون إبداء الأسباب.

يتكون الاستبيان من ثلاثة أجزاء. الجزء الأول يوثق بعض البيانات الشخصية عن المشارك, أما الجزء الثاني فيتكون من 11 سؤال مقالي قصير حول موضوع الدراسة. الجزء الثالث و الأخير هو جزء خاص لتوثيق أية ملاحظات إضافية قد يرغب المشارك في الإدلاء بها.

بعد الإجابة على كافة الأسئلة يرجى من المشاركين إرسال الاستبيان بالبريد الالكتروني كملف ملحق إلى العنوان التالي:
nasser.alhinai@yahoo.com

شاكرا لكم حسن تعاونكم.

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

Part I:

1. College *(Please select one):*

الكلية التي تنتمي إليها (اختر الإجابة المناسبة):

1.	2.	3.	4.	5.	6.	7.
Muscat مسقط	Shinas شناص	Musana المصنعة	Ibra ابراء	Nizwa نزوى	Salalah صلاله	Ibri عبري

2. Gender: 1. Male ( ) 2. Female ( )

الجنس ذكر أنثى

3. Your current English proficiency Level *(Please select one):*

مستواك الحالي في اللغة الانجليزية (اختر الإجابة المناسبة):

1.	2.	3.	4.
Pre-Elementary ما قبل المبتدئ	Elementary مبتدئ	Intermediate متوسط	Advanced متقدم

4. In which *Educational Region* did you receive all or most of your pre-college education? *(Please select one):*

في أي منطقة تعليمية تلقيت تعليمك العام أو الجزء الأكبر منه؟ (اختر الإجابة المناسبة):

1.	Muscat Governorate مسقط		5.	Southern Batinah Region جنوب الباطنة		9.	Al Wista Region الوسطى	
2.	Dhofar Governorate ظفار		6.	Northern Batinah Region شمال الباطنة		10.	Dakhiliya Region الداخلية	
3.	Musandam Governorate مسندم		7.	Southern Sharqiyah Region جنوب الشرقية		11.	Dhahira Region الظاهرة	
4.	Buraimi Governorate البريمي		8.	Northern Sharqiyah Region شمال الشرقية		12.	Other غير ذلك	

5. What type of school did you attend before joining college?*(Please select one):*

بأي فئة من المدارس التحقت قبل انضمامك للكلية؟ (اختر الإجابة المناسبة):

1.	Public school	مدرسة حكومية	
2.	Private school	مدرسة خاصة	
3.	Other	غير ذلك	

Part II:

Please answer the following questions in the space provided:

يرجى طباعة الإجابة على الأسئلة التالية في الأماكن المخصصة للإجابة تحت كل سؤال:

1. Before you joined the English Language Centre, what were your expectations about teachers and teaching in the college? Does what you have experienced so far meet your expectations?

ماذا كانت توقعاتك عن التدريس و المدرسين في الكلية قبل انضمامك إلى مركز اللغة الانجليزية ؟ و هل ما خبرته حتى الآن يطابق توقعاتك؟

2. Did you notice any differences in teaching style and methods between school teachers and college instructors? If YES, what are the differences?

هل لاحظت أية فروق في أنماط و طرق التدريس المتبعة بين معلم المدرسة و محاضر الكلية؟ في حال الإجابة بالإيجاب ما هي هذه الفروق؟

3. Which learning style do you prefer: teacher-cantered or independent study? And why?

ما هو نمط التعلم المفضل لديك: التعلم الموجه من قبل المحاضر أم التعلم الذاتي و المستقل؟ و لماذا؟

4. Have you been given the opportunity to evaluate the performance of your teachers in the Foundation Programme? Can you describe how this evaluation was conducted?
هل أعطيت لك الفرصة لتقييم أداء محاضريك في البرنامج التأسيسي؟ هل لك أن تصف كيف تمت عملية التقييم؟

5. In your opinion, should the Foundation Programme students be allowed to evaluate the teaching performance of their teachers? Why? Why not?
في رأيك هل يجب السماح لطلاب البرنامج التأسيسي بتقييم أداء محاضريهم؟ لماذا؟ لم لا؟

6. Do you think that the Foundation Programme students are capable of judging the teaching performance of their teachers? Why? Why not?
هل تعتقد بأن طلاب البرنامج التأسيسي لديهم القدرة على تقييم أداء محاضريهم؟ لماذا؟ لم لا؟

7. Do you think your evaluation strategies and criteria of teaching effectiveness have changed since you joined the college? If YES, in what way?
هل تعتقد بأن أسس و معايير الحكم على كفاءة التدريس لديك قد تغيرت منذ التحاقك بالكلية؟ إذا كانت الإجابة بالإيجاب كيف؟

8. In your opinion, what are the characteristics of effective college TESOL teachers?
في رأيك ما هي سمات محاضر اللغة الانجليزية الجيد و الكفء؟

9. Do you think that collecting student feedback will lead to improvements in teaching? Why do you think so?
هل تعتقد بأن إشراك الطالب في عملية تقييم أداء المدرس سوف ينتج عنه تحسن في جودة التدريس؟ لماذا في رأيك؟

10. Do you think that your evaluation of a teacher is affected by the views of your classmates?

هل تعتقد بأن تقييمك لأداء محاضريك يتأثر بآراء زملائك في الصف؟

11. Do you think that student evaluations of teaching should be considered in personnel decisions such as promotions and contract renewals? Why or why not?

هل تعتقد بأنه يجب أخذ تقييمات الطلاب لأداء محاضريهم بعين الاعتبار في اتخاذ القرارات الإدارية المتعلقة بترقيات و تجديد عقود المحاضرين؟ لماذا؟ لم لا؟

Part III: Any additional comments: *أية ملاحظات إضافية:*

Thank you for taking the time to complete this questionnaire

شكرا جزيلا لك على مشاركتك في هذه الدراسة

APPENDIX 6

Findings of the Qualitative Exploratory Study

Introduction

This summary is organised according to the three themes that emerged from the data: a) perceptions of the current staff appraisal practices in general, b) perceptions of the evaluation of teaching effectiveness and the characteristics of effective college teachers, and c) perceptions of students' role in the evaluation of college teaching. Once again, the first theme, perceptions of the staff appraisal system as a whole, is exclusively based on the data collected from administrators and lecturers only, as no questions probing this aspect were included in students' version of the questionnaire for the reasons stated under Section 5.2.2 in Chapter 5.

Perceptions of the Current Staff Appraisal Practices

Probably one of the most interesting findings about staff appraisal in the setting of this study is that program administrators generally tend to think of staff appraisal as once-a-year added *chore* rather than an important part of ongoing plans of staff professional development and/or quality assurance. Almost all of the program administrators surveyed quantified the time they spent in staff appraisal in terms of hours per lecturer per semester/ year. In addition, while all the program administrators surveyed consider quality control and assurance, maintaining accountability, and making personnel decisions as prime functions of staff appraisal in their departments, only around half of them seem to see the developmental potential of staff appraisal in their departments side by side with the evaluative one. As one GFP administrator puts it:

> *[Appraisal systems] are primarily for determining whether or not to continue to employ the teacher – it is evaluative rather than clinical supervision.*

In sharp contrast with the first, another administrator thinks that:

> *Developmental supervision is the key. Some teachers are at a level where they require accountability and quality control. Others are in need of less directive approaches. They are highly skilled practitioners who should be approached in a different manner.*

As for the lecturers' perceptions of the purpose of staff appraisal, almost half of the lecturers surveyed think that making personnel decisions and quality control are the prime purposes of staff appraisal in their GFP programs. Only three out of the 23 lecturers think that their appraisers are concerned with professional development and improvement in their departments. Furthermore, thirteen of the 23 lecturers were not exactly sure whether staff appraisal in their departments is evaluative, developmental, or a mixture of both, due to lack of transparency. Lecturers in general, however, expressed strong support for staff appraisal schemes which emphasise professional development and improvement as opposed to those which stress quality control and accountability. As one teacher wrote:

> *The evaluation process should be to improve the quality of the learning and teaching process, rather than being a "fault finding process" for non-professional reasons.*

Another major theme that emerged from the responses provided by the program administrators as well as the lecturers involved in this exploratory study is that classroom teaching is by far the most prominent factor considered in evaluating the overall performance of a lecturer. All of the program administrators and around two thirds of the lecturers surveyed think that classroom teaching constitutes the most important aspect of a lecturer's overall performance appraisal. However, while all the program administrators think that staff appraisal has improved the quality of teaching and/or learning in their departments, around one third of the lecturers think that staff appraisal has had no positive effect what so ever on their teaching or their students' learning. One lecturer describes the effect as:

> *None! If anything it made me less confident in my teaching ability. No constructive advice or feedback was ever offered. Only criticisms offered. In over ten years of teaching, this was the first negative comments I had received.*

Administrators and lecturers also disagree on the importance placed on personal attributes in staff appraisal. While slightly over one third of the administrators think that personal attributes is a major factor in the overall appraisal of a lecturer, around half of the lecturers think that the personal attributes of a lecturer constitute the second most important factor considered by program administrators in evaluating their lecturers. One lecturer thinks staff appraisal in his departments is

> *Mostly based on personal attributes. No objective assessment is made.*

Experience in the field also emerged as an important criterion used by program administrators in the overall evaluation of a lecturer. However, administrators seem to attach far more importance to this aspect than do lecturers. Six out of the eight administrators involved in this study considered lecturer's experience as an important factor to be taken into account in staff appraisal. On the other hand, only six lecturers out of 23 thought that their program administrators used experience as a factor in staff appraisal.

Team working skills and the ability and willingness of a lecturer to work effectively and professionally with other colleagues and participate in departmental committees and activities also featured as an important aspect of a lecturer's overall performance that is targeted by administrators in staff appraisal. Seven out of the eight GFP administrators surveyed regarded such skills as an important factor in staff appraisal. Lecturers, however, seem to be unaware of the importance their administrators attach to this aspect of their work. Only six out of 23 lecturers

mentioned team working skills and participation in committee work as a factor that is considered in staff appraisal.

Students' opinions and level of satisfaction with the overall performance of a lecturer was also pointed out by both administrators and lecturers as a factor taken into account in staff appraisal. However, while one fourth of the GFP administrators believe that students' opinions of the lecturer and their level of satisfaction with his/her teaching and overall performance is included in staff appraisal, only around one fifth of the lecturers think that their students' opinions and satisfaction form part of the staff appraisal scheme in their departments. One SET enthusiastic lecturer states:

> *Personally, I think the most important evaluation comes from students, and not from administrators or government officials.*

Another lecturer adds:

> *I think the effectiveness of our teaching can be assessed by the feedback from the students and their portfolio. Since students are very smart in this, they can immediately make out whether the teacher is capable or not.*

One fourth of the program administrators also regard a lecturer's commitment to the teaching profession and teaching duties as a factor that contributes to the overall evaluation of a lecturer's performance. Again, however, lecturers seem to place less importance on this aspect of their work compared to their supervisors. Only one fifth of the lecturers surveyed addressed this aspect, stressing on the importance of various forms of such commitment, such as systematic documentation of evidence and proper record keeping.

Both administrators and lecturers emphasised the effect of various cultural factors on the framing and implementation of staff appraisal in their departments. Among these factors is the cultural

association between staff appraisal and termination or accountability. One major weakness in staff appraisal as seen by an administrator is:

> *Over dependence on mandated formats designed for 'termination review'.*

Another GFP administrator adds:

> *Staff appraisal is always associated with termination.*

One of the lecturers describes her experience with the system:

> *My experience is that the evaluations were used in a destructive manner and carried out by people who were not qualified to do such evaluations in a professional manner.*

At the organisational level, factors such as the lack of transparency and poor management skills on the part of program and college administrators were also considered major factors shaping up the staff appraisal practices. As one lecturer describes it:

> *There may be an unwillingness to give direct, unambiguous information.*

Another lecturer adds:

> *Attitude of administration should be organizing rather being reckless, impulsive and punishing (terrorising).*

Students' obsession with grades rather than learning coupled with poor progression criteria for students from one level to another were also pointed out as sources of problems that may affect program administrators or students' evaluation of a lecturer's performance.

> *I think the desired outcomes are too optimistic given the level of the students. The situation with students is to ask for grades rather than earn them. Therefore, the students often move on to the next levels without full knowledge of the outcomes set by the Ministry of Manpower (for example, we have elementary level students in Technical Writing). This is a problem when teachers are evaluated and students are found to be beneath the level that they should be at.*

Perceptions of the evaluation of teaching effectiveness and the characteristics of effective college lecturers

Three main issues emerged from exploring this theme: a) perceptions of the evaluation of college teaching, b) perceptions of the characteristics of effective college teachers, and c) perceptions about the sources of data used in the evaluation of teaching effectiveness in the GFPs.

Perceptions of the evaluation of college teaching

It was found that slightly over half of the GFP administrators think that teaching is a profession which requires a lot of training and preparation. On the other hand, slightly over half of the lecturers surveyed believe that teaching is an art or partly an art which requires talent and 'passion'. One lecturer defines teaching:

> *I strongly think that it is an art. It is like any other sort of arts such as cooking which I really consider an art. When you teach you try different techniques and you use different materials that you select carefully in order to come up with the best and most desired students' outcomes. In cooking, the techniques are your cooking skills and the materials are the ingredients which all affect the final product which is the meal you want to make.*

Nevertheless, there seems to be strong consensus between the two groups on the need for evaluating the quality of teaching in higher education. Various mitigating reasons and concerns calling for proper teacher evaluation in the GFPs were expressed by both administrators as well as their lecturers. These concerns can be summed up as follows:

- the vast diversity in the educational and professional backgrounds of the lecturers;

  > *As staff come from different educational and professional backgrounds, we need to ensure that they fit to our situation and accordingly staff professional development programs can be based on the output of the evaluation and consequently improve their performance.*

- varying level/quality of educational qualifications and experience lecturers hold;

We now have instructors for the most part without formal teacher training. Trained teachers are required.

Here, as you know, teachers come with different levels of experience and education. Add to that the cultural diversity. Add to that the high turnover rate. It is difficult to make generalizations when the variables are so far ranging.

- the diverse cultural backgrounds of the lecturers in face of a very homogeneous student population;

Since most faculty members come from different educational and cultural backgrounds, it's imperative to evaluate their performance to insure that they comply with the system.

- poor recruitment practices which lack transparency and proper selection standards and procedures.

Some of the teachers recruited are joining colleges by chance or with some other motives and are not basically interested in teaching.

Teachers should be "evaluated" during the hiring process. After that, they should be treated as responsible, professional teachers. They should not be monitored constantly as that only serves to degrade them and reduce efficiency and quality of performance.

Students' understanding, perceptions, and expectations about teaching at college level emerging from this exploratory study seem to add to the complexity of the contextual circumstances mentioned above. The students sampled in this study explicitly expressed stronger preference for teacher-centred teaching. Thirteen out of the 20 students surveyed in this study prefer teacher-centred approaches in teaching at college. One student says:

Student-centred approaches may be good, but I think they are a waste of student's time. Trying to understand a subject on my own may take me 30 minutes, while the teacher can explain the whole thing in 10 or 15 minutes. The latter is better!

Another student adds:

> *Having a teacher with good knowledge of the subject matter in the classroom is very important. The teacher should be able to give the students all the right information they need, especially information which is not included in the textbook.*

A third student points out:

> *[I prefer] teacher-centred. To be honest, I do not think I'll learn anything if I was given a choice to learn by myself.*

Only four out of the 20 students expected teaching styles and techniques at college to be different from those used in schools and twelve students out of 20 think that their criteria for judging the quality of teaching in college classrooms are still the same as those they used in primary and secondary schools. Three students out of 20 thought that English classes at college are taught only by native speakers of English.

> *I expected all lecturers to be native speakers of English, so that they can teach us their mother tongue.*

Almost one third of the students surveyed also expect college teachers to be more caring and supportive to students than school teachers.

Perceptions of the characteristics of effective college teachers

Analysis of the administrators', lecturers', and students' perceptions of the characteristics of effective college TESOL lecturers revealed varying degrees of mismatch between the views of the three groups. While differences between the administrators' and lecturers' perceptions of what constitutes effective college teaching are mostly in ranking of importance, the differences between the administrators' and lecturers' perceptions of effective TESOL lecturers on one side and students' perceptions of the same on the other side seem to represent a shift in priorities. These perceptions are summarised in the table that follows for the three sub-groups. For each

group, the list of characteristics of effective college TESOL lecturers is ranked by frequency (in brackets), from the most frequent to the least frequent.

As can be seen in the table below, five characteristics of effective TESOL lecturers are highlighted by administrators, lecturers, and students alike despite the differences in ranking of importance attached to these characteristics by three groups. These are: knowledge of the subject matter, dedication and passion for teaching, having sufficient teaching experience, flexibility and diversity in teaching style, and having strong personality and good classroom management skills.

Perceptions of the characteristics of effective college TESOL lecturers

Administrators (N= 8)	**Lecturers (N= 23)**	**Students (N=20)**
Academic qualifications and formal teacher training (4) *	Academic qualifications and formal teacher training (11)*	Clear pronunciation and accent (15)**
Teaching experience (3)*	Showing care and support for students (11)*	Respect for students (8)**
Knowledge of the subject matter (3)*	Knowledge of the subject matter (10)*	Showing care and support for students (8)*
Dedication and passion for teaching (3)*	Flexibility and diversity in teaching style (10)*	The ability to give clear explanations (7)**
The ability to promote independent learning (3)*	Dedication and passion for teaching (8)*	The ability to make classes interesting (5)**
The ability to encourage and motivate students (3)**	Communication skills (6)*	Knowledge of the subject matter (4)*
Strong personality (2)*	Teaching experience (5)*	Dedication and passion for teaching (4)*
Showing sensitivity to the culture and system in place (2)**	Organisation and planning skills (4)*	Teaching experience (3)*
Flexibility and diversity in teaching style (2)*	The ability to promote independent learning (3)*	Flexibility and diversity in teaching style (3)*
Communication skills (2)*	Making good use of instructional aids (3)**	Strong personality and good classroom management (3)*
Organisation and planning skills (2)*	Strong personality (3)*	
Team working skills (1)**		

*** Shared with other groups**
**** Unique to this group**

Looking at the top 5 ranked characteristics of effective teaching from the perspective of the three groups, some interesting differences can be observed. While four of top 5 characteristics of effective teaching from both the administrators'' and teachers' perspectives have matches in each other's lists, four of the top 5 rated characteristics of effective teachers on the students' list have no matches on either the administrators'' or the teachers' criteria. Both administrators and lecturers consider having proper academic qualifications as the most important characteristic of an effective college TESOL lecturer. Both also seem to attach great importance to subject mastery and dedication to teaching.

When looking at the administrators'' and lecturers' lists as a whole and not only the top 5 rated traits, it can be seen that administrators appeared to stress on the importance of three characteristics that had no direct match on the lecturers' list of priorities. Features of effective teaching such as the ability to encourage and motivate students, showing sensitivity to the culture and system in place, and team working skills were highlighted by the administrators only but not by their lecturers. On the other hand, two attributes of effective teaching, namely lecturer's ability to show care and support for students and making good use of instructional aids, seem to be more of a concern for lecturers but, surprisingly, not for their administrators.

As pointed out above, students in the GFP seem to have different priorities in evaluating teaching effectiveness despite the few similarities with their lecturers and program administrators. Four of students' top ranked characteristics of effective teaching, namely: having clear pronunciation and accent, respect for students, the ability to give clear explanations, and the ability to make classes interesting, have no direct matches on administrators' and lecturers lists of characteristics of good teachers. Furthermore, some dimensions of teaching effectiveness that were stressed by both

administrators and lecturers, such as having good organisation and planning skills, and the ability to promote independent learning, were completely overlooked by the students.

Two mismatches in perceptions of effective teaching between students on one side and lecturers and administrators on the other may be particularly important. These concern the importance of having clear pronunciation and accent and the ability to promote independent learning. While having clear pronunciation is a basic requirement in verbal communication, especially in teaching situations where the medium of instruction is a foreign or a second language, it is not clear what students mean by 'clear accent'. It is not also clear whether students realise the difference between accent and intonation and stress. The distinction here is very important, especially in the context of the GFPs where lecturers are multinational and come from diverse educational and cultural backgrounds. Confusing the two concepts may lead students to have preferences for a certain group of teachers and be biased against other groups irrespective of their teaching performance in the classroom.

Disagreement on the importance of independent learning also emerges as an important theme. The implications of this mismatch, between students on one side and lecturers and administrators on the other, regarding their perceptions of the importance of independent learning and student-centred approaches of teaching can be far-reaching and demoralising for teachers. In a situation where only administrators and lecturers, but not their students, see the potential of student-centred approaches in effective teaching, designing teaching evaluation tools for this dimension of teaching in the classroom would require careful consideration and study. Failure to recognise the effect of such mismatches in priorities may lead to independent learning advocates being punished by their students, and possibly by their administrators where students' ratings are collected, for all the wrong reasons.

Perceptions about the sources of data used in the evaluation of teaching effectiveness in the GFPs

The findings about the sources of information program administrators use to evaluate teaching in their departments confirmed in part the researcher's expectations. The most common source of data used by GFP administrators in evaluating the teaching performance of their lecturers is classroom observation and visitation. All of the administrators and seventeen out of the 23 lecturers listed class visitation as the number one source of data on teaching evaluation. Students' feedback, which was used by both administrators and lecturers to incorporate informal oral feedback, students' complaints, and unsystematic students' ratings, was found to be the second most used source of data as judged by both administrators and lecturers. As one GFP administrator puts it:

> *Direct observation is used primarily. Although student feedback and colleague feedback provides very biased information it does give a picture of how well the instructor is perceived. Required but is not an effective means of viewing competencies. Used as a 'satisfaction rating'.*

However, while three quarters of the administrators said that they used students' feedback in evaluating their teachers, only about half of the lecturers believe that students' feedback is used in evaluating their teaching. One lecturer explains why:

> *I don't think they are ready for it yet. For any evaluation, the most important thing needed is an objective view of things. This cannot be expected from students at this level because they are here not on a motivation to develop themselves. Most of them happen to be here and they have not been guided as to the real reason for their presence here. Now this has lead to all the present-day problems of indiscipline on campus. The only thing that the vast majority of them want is the marks - by hook or crook as only a small percentage is actually working for them! The cheating rate has shot up!*

Commenting on students' ability to rate their teachers' performance, another lecturer adds:

> *They are up to a point, but some opinions given are misleading. For example our students are asked if we start and end class on time. One student said his teacher started TOO EARLY! Meaning she started on time but marked him late, which of course, he didn't like. Also in most classes students will say teachers are suitably dressed, but there are always one or two who object to what teachers wear and you never quite know why. The teachers they comment on are usually correctly dressed for a work environment. As to teaching performance, sometimes they say things are difficult or boring when I think they really mean it requires a bit of effort. Sometimes students just don't understand our techniques because of the educational background they have come from.*

The third most used data source is peer opinion. Again, there seems to be discrepancy between what administrators say and what lecturers see in practice. About two thirds of the GFP administrators point out peer opinion as a source of information they use in making judgments about the teaching performance of their lecturers. However, only around one third of the lecturers agree with this.

The use of two other sources of data, namely self assessment and students' achievement, also seem to be an area of disagreement between the two groups. One fourth of the administrators claim that lecturers are offered the chance to carry out self assessment as part of the teaching evaluation process. However, none of the lecturers confirmed this. Instead, some lecturers pointed out that students' achievement in tests is being used by program administrators in forming inferences about the quality of teaching in their departments, a practice that was not mentioned by the administrators at all.

Such discrepancies could point to serious problems in communication between GFP administrators and their lecturers or lack of transparency in policymaking and implementation in the institution governing these programs as a whole. Such concerns are amplified by the fact that slightly over one third of the lecturers surveyed do not know for sure whether there is a system in place in their programs for teaching evaluation or not. One lecturer pointed out:

> *I don't think it is developmental at all, let alone being evaluative. I still see inefficient teachers who ought to be sent away hanging around. This not only damages the young mind of the student community, but also makes it difficult for other teachers to bring these students on track after the "damage"!*

Another said:

> *Not sure as we are not given any feedback after the evaluation. It seems like it is a routine evaluation.*

Administrators' subjectivity, bias, lack of transparency, and poor communication with the different stakeholders were concerns that surfaced frequently in lecturers' responses.

Perceptions of students' role in the evaluation of teaching effectiveness

It became evident form the findings of this study that systematic SET is still a novelty in the programs surveyed. Students' role in the evaluation of their lecturer's teaching performance is minimum in some colleges and almost nonexistent in others. In colleges where students' input is sought, the process remains largely informal and students' feedback or 'complaints' are usually taken verbally and when required by the head of department.

Only three out of the eight GFP administrators reported using students' surveys. Even when students' ratings are collected using a questionnaire, the process itself remains unsystematic and lacks consistency and clear regulations. In the college where it was reported by the GFP administrators and some lecturers that SETs were collected using a questionnaire, slightly less than half of the students reported that they had never been asked to rate their teachers in the college. This is probably because the administration of the rating surveys is selective and does not include all classes or teachers. As two of the administrators stated, students' evaluations are only considered when the department head's evaluation of the teacher is poor.

[Students play] a marginal role only. Students' feedback is sometimes coloured by non-academic factors and considerations. But when a teacher's performance rating is poor in the administration's evaluation and the student feedback, the assessment made by the students will be considered as important.

Overall, only about one fourth of the teachers surveyed said that students in their departments were given the chance to rate their teachers using surveys. Slightly over one third of the lecturers do not know for sure what role their students play in the evaluation of teaching.

The almost unnoticeable role students play in evaluating teaching effectiveness in the GFPs in practice is probably a direct result of the general view of suspicion and distrust held by many administrators and lecturers towards students' ratings as evident from their responses in this study. While some program leaders and lecturers surveyed had mixed feelings about involving students in the evaluation of teaching, the general predisposition towards SET among the majority of the two groups was marked by suspicion and distrust. GFP administrators' and lecturers' reservations towards SET can be summed up as follows:

- Students are not capable of judging the various 'trade secrets' and complex dimensions of teaching, such as lecturer's linguistic ability and choice of teaching methods.

They can be expected to comment on their satisfaction with the instruction. They can not be expected to judge the 'trade skills' of the instructor.

- Students' objectivity is susceptible to various extraneous factors irrelevant to teaching and, therefore, students' ratings should be viewed with caution.

Unfortunately, I do not believe students in the Foundation Program are capable of evaluating teachers as they are not interested in their education. I don't think that these students are interested in actually learning something but rather just want to be passed the whole way and given a degree simply for attending class. Asking a student who doesn't care about their education what he or she thinks about their teacher's performance isn't a good way of conducting evaluations because the information will be biased.

- The quality and usefulness of students' evaluations are dependent on student's maturity and understanding of the teaching/learning situation in the college.

  Since the students themselves are not motivated in their studies and are not properly oriented in learning, they cannot assess the teacher's professional abilities.

- Like any other qualified assessor or observer of teaching, students should be trained on the dimensions of college teaching before they are allowed to rate their lecturers.

  They are capable provided that they receive proper instructions and guidance. Sometimes, their objectivity is questionable.

  I feel that the students need to be informed before asking for their assessment. They have to understand the seriousness of the assessment – that it is to help the teacher to improve teaching strategies etc.

Students, however, have a different view to the above. Seventeen out of the twenty students involved in this study stated that students in the GFPs are capable of judging the teaching performance of their lecturers and, therefore, should be allowed to evaluate teaching. One student says:

> *This is a very important issue. Students are the most knowledgeable of their teacher's effectiveness. This is because students are in constant contact with their teachers and they know what happens in the classroom more than abybody else. Through students' ratings one can measure teachers' competence in teaching, especially in the Foundation program where teacher's role is very important.*

Students also seem to appreciate the confidentiality element in systematic SETs compared to other less formal means of obtaining students' feedback, such as verbal complaints. Several students pointed out that students usually refrain from complaining to the administration about the performance of a teacher even if there is a major cause of concern for fear that their identities will be revealed to the teacher. One student explains:

> *... Some students don't like the way the teacher treats them or the way he teaches them and they can't talk [openly]. So the best way is to write what they feel, what they like and what they don't….*

The majority of students also believe that they can give objective evaluation of their teachers' performance in the classroom without being affected by their classmates' views. Most of them also believe that collecting students' feedback will result in improvement in teaching practices. As one student points out:

> *Students' evaluation of teaching, especially those collected from the Intermediate and Advanced levels, can provide administrators with valuable indicators on the effectiveness of teaching in the program and the effectiveness of the program as a whole in meeting its objectives.*

Another student cautions against the exclusion of students from the evaluation of teaching in the GFP and argues strongly that:

> *If students are not given the chance to evaluate the performance of their teachers, the quality of teaching will suffer and teachers may become careless and indifferent to their students' problems on the basis that students have no voice in what happens in the classroom.*

Students' opinion is divided, however, over the use of SET data in making personnel decisions, such as promotions and contract renewal. Only about half of the students surveyed support using the results of students' ratings in making personnel decisions. The rest of the students believe SET data should not be used for this purpose or used only if considerable care is taken in collecting the data and interpreting the results.

Summary

This exploratory qualitative email survey experimented with a relatively new method of data collection that is rarely used in the context of the study. The method proved to be very time and cost efficient. The main objective of this study was to explore the subject of staff appraisal and the evaluation of teaching effectiveness as practiced and perceived in the context of the GFPs in colleges of technology in Oman. The study was not meant to draw conclusive evidence or make

generalisations about the subject or the participants, but rather to help generate qualitative data that can be examined for patterns or themes that merit further larger scale investigation. It was also hoped that the data gathered would help in the development of the research instruments for the main study.

Two of the most important themes that emerged from the data of this study are:

1. The mismatch between the administrators and lecturers on one side and students on the other in their perceptions of the characteristics of effective college teachers were unexpected and revealed an important difference in priorities which may potentially have serious implications on teaching and the evaluation of teachers in colleges of technology in Oman.
2. The mismatches between lecturers' and students' perceptions of students' evaluation of teaching and the role of students as evaluators of college teaching were also significant. Unlike students who were very enthusiastic for SET, teachers and administrators in general seemed to distrust students and their ratings of teaching on the basis that students are 'not ready' for this yet or because their ratings are 'unreliable' and often 'coloured by non-academic considerations'.

These two findings were turning points in the research project. The two themes will be the backbone of the main larger scale quantitative study that will follow. Integrating the characteristics of effective lecturers identified by the respondents in this study with those identified in the prevailing research literature in the subject, the study that follows will investigate the matches and mismatches in priorities between lecturers and students mentioned above further, making use of a much bigger and more representative sample of the population. The same procedure will also be replicated with the second theme, perceptions of SET and the role of the

student in evaluating teaching. As part of this latter theme also, the main study will administer a widely used American standardised SET form to assess students' ability to produce reliable evaluations of their lecturers and to determine whether students are capable of identifying the teaching dimensions underlying this instrument.

APPENDIX 7

May 2008

Dear students,

Thank you for agreeing to complete this questionnaire. The main purpose of this study is to evaluate the use of students' evaluation of college teaching for monitoring and improving teaching effectiveness and for professional development in the Foundation Programme in the Colleges of Technology in Oman. The research also investigates various factors that may affect students' ratings of their teachers in the context of the Foundation Programme and ways to improve the reliability, validity, and utility of students' evaluation of teaching effectiveness. The study, which forms part of my doctoral programme, is necessary to provide reliable information to decision makers and educators on how student ratings can be used to promote quality in the Foundation Programmes in Colleges of Technology.

This questionnaire is one of two questionnaires used in this study. Its main purpose is to examine how Foundation Programme administrators, lecturers, and students perceive good college teaching. It also investigates their views about students' evaluations/ratings of their lecturers.

Your participation in this study is voluntary. You may withdraw from the study at any time for any reason and without prejudice.

All data will be treated confidentially. Information obtained about you and the views you express in your answers will not be shared with your College; neither will your identity be disclosed in the research report.

Please answer all the questions following the instructions given at the beginning of each section.

Once more, thank you for participating in this study.

Sincerely,

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

PERCEPTIONS OF 'GOOD' COLLEGE TEACHING AND STUDENT EVALUATION OF TEACHERS

(STUDENT'S VERSION)

Student ID: _____ College: _____ Group No.: ____
English Proficiency Level: _____ Gender: _____ Date: _____

Section One: Perceptions Of 'Good' College Teaching

Below is a list of characteristics that are considered by many to be important to good teaching and effective college teachers. Please rate the importance of these characteristics from your point of view, and in relation to teaching English as a second/foreign language to adults at college level. Remember that there is no correct or wrong answer. For each statement, simply indicate the response closest to your view by circling the appropriate number on the scale provided. Please answer all the questions.

☞	❶	❷	❸	❹	❺
	Not at all important	**Slightly important**	**Moderately important**	**Very important**	**Extremely important**

1. Having full command of the subject matter	1	2	3	4	5
2. Showing strong enthusiasm for the subject	1	2	3	4	5
3. Making the course intellectually challenging and stimulating	1	2	3	4	5
4. Being able to stimulate the interest of the students in the subject	1	2	3	4	5
5. Showing dedication to teaching	1	2	3	4	5
6. Preparing good course materials and carefully explaining them to students	1	2	3	4	5
7. Making students feel welcome in seeking help/advice in or outside of class	1	2	3	4	5
8. Having relevant Academic qualifications in teaching English as a second/foreign language	1	2	3	4	5
9. Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject	1	2	3	4	5
10. Presenting points of view other than lecturer's own when appropriate	1	2	3	4	5
11. Using lively presentation styles which hold students' interest during class	1	2	3	4	5
12. Making proper use of instructional media, teaching aids, and multi-media labs	1	2	3	4	5
13. Demonstrating full compliance with the announced objectives of the course	1	2	3	4	5
14. Being dynamic and energetic in conducting the class	1	2	3	4	5
15. Encouraging students to participate in classroom activities and discussions	1	2	3	4	5
16. Having a genuine interest in individual students	1	2	3	4	5

❶	❷	❸	❹	❺
Not at all important	**Slightly important**	**Moderately important**	**Very important**	**Extremely important**

17. Presenting the background or origin of ideas/concepts developed in class	1	2	3	4	5
18. Having sufficient formal teacher training in teaching English as a second/foreign language	1	2	3	4	5
19. Keeping abreast of the latest developments in the field or subject	1	2	3	4	5
20. Inviting students to share their ideas and knowledge	1	2	3	4	5
21. Demonstrating good skills in classroom management	1	2	3	4	5
22. Showing respect for all students	1	2	3	4	5
23. Being available to students for advice and support in or after class	1	2	3	4	5
24. Showing sensitivity to the culture of the organisation and society at large	1	2	3	4	5
25. Demonstrating very good use of student-centred approaches	1	2	3	4	5
26. Encouraging students to ask questions and ensuring that answers given to students are meaningful	1	2	3	4	5
27. Showing flexibility and diversity in teaching style	1	2	3	4	5
28. Enhancing presentation with the use of humour	1	2	3	4	5
29. Having native-like intonation and stress	1	2	3	4	5
30. Demonstrating a high level of expressiveness and giving clear explanations	1	2	3	4	5
31. Giving assignments/graded materials which test class content as emphasised by the lecturer	1	2	3	4	5
32. Encouraging students to express their own ideas and/or question the lecturer	1	2	3	4	5
33. Having relevant and sufficient experience in teaching English as a second/foreign language	1	2	3	4	5
34. Being friendly toward individual students	1	2	3	4	5
35. Giving valuable feedback on assessments/graded material	1	2	3	4	5
36. Giving lectures/tutorials in a style and pace that facilitate note-taking	1	2	3	4	5
37. Being a native speaker of the target language	1	2	3	4	5
38. Using appropriate and fair methods of evaluating student work	1	2	3	4	5

Section Two: Perceptions of Student Evaluation of Teachers

Student evaluation/rating of teaching is one of the most commonly used teaching evaluation methods in universities and colleges worldwide. This section of the questionnaire explores <u>your views and perceptions of student evaluations/ratings of college teachers</u>. For each statement, please indicate the response closest to your view by circling the appropriate number on the scale provided.

☞	❶	❷	❸	❹
	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**

39. Student ratings can provide reliable and valid diagnostic feedback to lecturers for improving teaching	1	2	3	4
40. Student ratings of lecturers should be used as one source of data for making personnel decisions such as lecturers' contract renewal or termination	1	2	3	4
41. Student ratings of teaching can provide good protection to lecturers against potential biases in evaluation by heads of department	1	2	3	4
42. Student ratings of lecturers should only be used for teaching improvement purposes, not to hold lecturers accountable for deficiencies in their performance	1	2	3	4
43. Students in the Foundation Programme are capable of rating most aspects of a lecturer's teaching performance	1	2	3	4
44. A student's prior interest in the subject affects his/her rating of this subject's instructor	1	2	3	4
45. Students give lower ratings to teachers of courses with high workload	1	2	3	4
46. Involving students and lecturers in developing rating forms will create a common ground for both parties to develop shared meanings of 'good' college teaching	1	2	3	4
47. Lecturer's personal attributes play a major role in students' ratings	1	2	3	4
48. Creative college lecturers may be poorly rated by students if these lecturers are unaware of the learning styles students are used to in pre-college education	1	2	3	4
49. Students should be trained in using rating forms and told what each item represents before they are asked to rate their lecturers	1	2	3	4
50. I feel that students' rating of/reaction to my teaching is significantly influenced by the fact that I am a native/non-native speaker of English	1	2	3	4
51. Student evaluation of teaching can cause lecturers to deflate course workload and lower standards in order to keep students happy	1	2	3	4
52. The grade students expect in a course affects how they rate their lecturer in that course	1	2	3	4
53. Students' ratings of lecturers for making personnel decisions such as contract renewal or termination is a main cause of grade inflation	1	2	3	4
54. Lecturers are more likely to receive high students' ratings when evaluated by students of the opposite sex	1	2	3	4
55. Lecturers with good communication skills are more likely to receive high students' ratings	1	2	3	4
56. Educating students about the generic characteristics of effective college teaching will improve the reliability and validity of their ratings	1	2	3	4
57. Students' judgment of the teaching performance of a lecturer can be affected by the lecturer's ethnic background or nationality	1	2	3	4

Thank you for taking the time to complete this questionnaire

APPENDIX 8

May 2008

Dear Colleague,

Thank you for agreeing to complete this questionnaire. The main purpose of this study is to evaluate the use of students' evaluation of college teaching for monitoring and improving teaching effectiveness and for professional development in the Foundation Programme in the Colleges of Technology in Oman. The research also investigates various factors that may affect students' ratings of their teachers in the context of the Foundation Programme and ways to improve the reliability, validity, and utility of students' evaluation of teaching effectiveness. The study, which forms part of my doctoral programme, is necessary to provide reliable information to decision makers and educators on how student ratings can be used to promote quality in the Foundation Programmes in Colleges of Technology.

This questionnaire is one of two questionnaires used in this study. Its main purpose is to examine how Foundation Programme administrators, lecturers, and students perceive good college teaching. It also investigates their views about students' evaluations/ratings of their lecturers. The questionnaire consists of three sections. **Section 1** surveys participants' perceptions of good college teaching. **Section 2** explores their views about students' evaluations/ratings of college lecturers. **Section 3** asks for background information about the participant.

Your participation in this study is voluntary. You may withdraw from the study at any time for any reason and without prejudice.

All data will be treated confidentially. Information obtained about you and the views you express in your answers will not be shared with your College; neither will your identity be disclosed in the research report.

Completed questionnaires should be returned directly to the researcher who will be available in your department during the administration of the questionnaire for any advice or guidance you may require in completing the questionnaire.

Once more, thank you for participating in this study.

Sincerely,

Nasser Al Hinai
PhD researcher
School of Education
University of Durham
United Kingdom

PERCEPTIONS OF 'GOOD' COLLEGE TEACHING AND STUDENT EVALUATION OF TEACHERS

(TEACHER'S VERSION)

Section One: Perceptions Of 'Good' College Teaching

Below is a list of characteristics that are considered by many to be important to good teaching and effective college teachers. Please rate the importance of these characteristics from your point of view, and in relation to teaching English as a second/foreign language to adults at college level. Remember that there is no correct or wrong answer. For each statement, simply indicate the response closest to your view by circling the appropriate number on the scale provided. Please answer all the questions.

☞	❶ Not at all important	❷ Slightly important	❸ Moderately important	❹ Very important	❺ Extremely important				
1. Having full command of the subject matter					1	2	3	4	5
2. Showing strong enthusiasm for the subject					1	2	3	4	5
3. Making the course intellectually challenging and stimulating					1	2	3	4	5
4. Being able to stimulate the interest of the students in the subject					1	2	3	4	5
5. Showing dedication to teaching					1	2	3	4	5
6. Preparing good course materials and carefully explaining them to students					1	2	3	4	5
7. Making students feel welcome in seeking help/advice in or outside of class					1	2	3	4	5
8. Having relevant Academic qualifications in teaching English as a second/foreign language					1	2	3	4	5
9. Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject					1	2	3	4	5
10. Presenting points of view other than lecturer's own when appropriate					1	2	3	4	5
11. Using lively presentation styles which hold students' interest during class					1	2	3	4	5
12. Making proper use of instructional media, teaching aids, and multi-media labs					1	2	3	4	5
13. Demonstrating full compliance with the announced objectives of the course					1	2	3	4	5
14. Being dynamic and energetic in conducting the class					1	2	3	4	5
15. Encouraging students to participate in classroom activities and discussions					1	2	3	4	5
16. Having a genuine interest in individual students					1	2	3	4	5

(Continued next page)

❶	❷	❸	❹	❺
Not at all important	**Slightly important**	**Moderately important**	**Very important**	**Extremely important**

17. Presenting the background or origin of ideas/concepts developed in class	1	2	3	4	5
18. Having sufficient formal teacher training in teaching English as a second/foreign language	1	2	3	4	5
19. Keeping abreast of the latest developments in the field or subject	1	2	3	4	5
20. Inviting students to share their ideas and knowledge	1	2	3	4	5
21. Demonstrating good skills in classroom management	1	2	3	4	5
22. Showing respect for all students	1	2	3	4	5
23. Being available to students for advice and support in or after class	1	2	3	4	5
24. Showing sensitivity to the culture of the organisation and society at large	1	2	3	4	5
25. Demonstrating very good use of student-centred approaches	1	2	3	4	5
26. Encouraging students to ask questions and ensuring that answers given to students are meaningful	1	2	3	4	5
27. Showing flexibility and diversity in teaching style	1	2	3	4	5
28. Enhancing presentation with the use of humour	1	2	3	4	5
29. Having native-like intonation and stress	1	2	3	4	5
30. Demonstrating a high level of expressiveness and giving clear explanations	1	2	3	4	5
31. Giving assignments/graded materials which test class content as emphasised by the lecturer	1	2	3	4	5
32. Encouraging students to express their own ideas and/or question the lecturer	1	2	3	4	5
33. Having relevant and sufficient experience in teaching English as a second/foreign language	1	2	3	4	5
34. Being friendly toward individual students	1	2	3	4	5
35. Giving valuable feedback on assessments/graded material	1	2	3	4	5
36. Giving lectures/tutorials in a style and pace that facilitate note-taking	1	2	3	4	5
37. Being a native speaker of the target language	1	2	3	4	5
38. Using appropriate and fair methods of evaluating student work	1	2	3	4	5

Section Two: Perceptions of Student Evaluation of Teachers

Student evaluation/rating of teaching is one of the most commonly used teaching evaluation methods in universities and colleges worldwide. This section of the questionnaire explores <u>your views and perceptions of student evaluations/ratings of college teachers</u>. For each statement, please indicate the response closest to your view by circling the appropriate number on the scale provided.

☞	❶	❷	❸	❹
	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**

39. Student ratings can provide reliable and valid diagnostic feedback to lecturers for improving teaching	1	2	3	4
40. Student ratings of lecturers should be used as one source of data for making personnel decisions such as lecturers' contract renewal or termination	1	2	3	4
41. Student ratings of teaching can provide good protection to lecturers against potential biases in evaluation by heads of department	1	2	3	4
42. Student ratings of lecturers should only be used for teaching improvement purposes, not to hold lecturers accountable for deficiencies in their performance	1	2	3	4
43. Students in the Foundation Programme are capable of rating most aspects of a lecturer's teaching performance	1	2	3	4
44. A student's prior interest in the subject affects his/her rating of this subject's instructor	1	2	3	4
45. Students give lower ratings to teachers of courses with high workload	1	2	3	4
46. Involving students and lecturers in developing rating forms will create a common ground for both parties to develop shared meanings of 'good' college teaching	1	2	3	4
47. Lecturer's personal attributes play a major role in students' ratings	1	2	3	4
48. Creative college lecturers may be poorly rated by students if these lecturers are unaware of the learning styles students are used to in pre-college education	1	2	3	4
49. Students should be trained in using rating forms and told what each item represents before they are asked to rate their lecturers	1	2	3	4
50. I feel that students' rating of/reaction to my teaching is significantly influenced by the fact that I am a native/non-native speaker of English	1	2	3	4
51. Student evaluation of teaching can cause lecturers to deflate course workload and lower standards in order to keep students happy	1	2	3	4
52. The grade students expect in a course affects how they rate their lecturer in that course	1	2	3	4
53. Students' ratings of lecturers for making personnel decisions such as contract renewal or termination is a main cause of grade inflation	1	2	3	4
54. Lecturers are more likely to receive high students' ratings when evaluated by students of the opposite sex	1	2	3	4
55. Lecturers with good communication skills are more likely to receive high students' ratings	1	2	3	4
56. Educating students about the generic characteristics of effective college teaching will improve the reliability and validity of their ratings	1	2	3	4
57. Students' judgment of the teaching performance of a lecturer can be affected by the lecturer's ethnic background or nationality	1	2	3	4

Section Three: Background Information

✍ **Please tick (✓) as applicable:**

58. College:

1.	2.	3.	4.	5.	6.	7.
Muscat	Shinas	Musana	Ibra	Nizwa	Salalah	Ibri

59. Gender:

1.	Male	
2.	Female	

60. Age group:

1.	2.	3.	4.	5.	6.	7.	8.
Under 25	25-30	31-35	36-40	41-45	46-50	51-60	Over 60

61. Position/ Job title:

1.	Director of English Language Centre	
2.	Head of Section	
3.	Lecturer	

62. What is your mother tongue?

1.	2.	3.	4.	5.	6.
Arabic	English	Urdu	Hindi	French	Other: (*please specify)*
					……..

63. Years of experience in English language teaching:

1.	2.	3.	4.	5.
0-5 years	6-10 years	11-15 years	16-20 years	More than 20 years

64. Ethnic background:

1.	2.	3.	4.	5.	6.	7.
Omani Arab	Non-Omani Arab	South-western Asian	European	African	North American	Other: (*please specify)*
						…….

Thank you for taking the time to complete this questionnaire

APPENDIX 9

STUDENTS' EVALUATION OF EDUCATIONAL QUALITY
The Foundation Programme
Colleges of Technology- Oman

Student's ID No.: …… College: …… Group No.: ……
Lecturer's Name: …… Course: …… …… Student's Gender: ……
Student's Current English Proficiency Level:…… Date: ……

This questionnaire gives you an opportunity to express anonymously your views about the teaching effectiveness of your lecturer in this course. Please indicate the response closest to your view by circling the appropriate number in front of each statement.

☞ ❶	❷	❸	❹	❺
Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**

LEARNING AND ACADEMIC VALUE					
1. You are finding the course intellectually challenging and stimulating	1	2	3	4	5
2. You are learning something which you consider valuable	1	2	3	4	5
3. Your interest in the subject is increasing as a consequence of this course	1	2	3	4	5
4. You are learning and understanding the subject materials of this course	1	2	3	4	5
LECTURER ENTHUSIASM					
5. Lecturer is enthusiastic about teaching the course	1	2	3	4	5
6. Lecturer is dynamic and energetic in conducting the course	1	2	3	4	5
7. Lecturer enhances presentation with the use of humour	1	2	3	4	5
8. Lecturer's style of presentation holds your interest during class	1	2	3	4	5
ORGANIZATION/CLARITY					
9. Lecturer's explanations are clear	1	2	3	4	5
10. Course materials are well prepared and carefully explained	1	2	3	4	5
11. Teaching/learning activities fit in with the announced course objectives	1	2	3	4	5
12. Lecturer gives lectures/tutorials that facilitate taking notes	1	2	3	4	5
GROUP INTERACTION					
13. Students are encouraged to participate in class discussions	1	2	3	4	5
14. Students are invited to share their ideas and knowledge	1	2	3	4	5
15. Students are encouraged to ask questions and are given meaningful answers	1	2	3	4	5
16. Students are encouraged to express their own ideas and/or question the lecturer	1	2	3	4	5
INDIVIDUAL RAPPORT					
17. Lecturer is friendly toward individual students	1	2	3	4	5
18. Lecturer makes students feel welcome in seeking help/advice in or outside of class	1	2	3	4	5
19. Lecturer has a genuine interest in individual students	1	2	3	4	5
20. Lecturer is adequately accessible to students during office hours or after class	1	2	3	4	5
BREADTH OF COVERAGE					
21. Lecturer contrasts the implications of various theories	1	2	3	4	5
22. Lecturer presents the background or origin of ideas/concepts developed in class	1	2	3	4	5
23. Lecturer presents points of view other than his/her own when appropriate	1	2	3	4	5
24. Lecturer adequately discusses current developments in the field	1	2	3	4	5

ASSESSMENT/GRADING					
25. Feedback on assessments/graded material is valuable	1	2	3	4	5
26. Methods of evaluating student work are fair and appropriate	1	2	3	4	5
27. Assessments/graded materials test class content as emphasized by the lecturer	1	2	3	4	5
ASSIGNMENTS/READINGS					
28. Required readings/texts are valuable	1	2	3	4	5
29. Readings, homework, etc. contribute to appreciation and understanding of the subject	1	2	3	4	5
OVERALL RATING **(1= Very poor…2= Poor…3= Average…4= Good…5= Very Good)**					
30. Overall, how does this class compare with other classes in the Programme?	1	2	3	4	5
31. Overall, how does this lecturer compare with other lecturers in the Programme?	1	2	3	4	5

BACKGROUND COURSE/CLASS CHARACTERISTICS
32. Course difficulty, relative to other courses, is: 1. Very Easy 2. Easy 3. Medium 4. Hard 5. Very Hard
33. Course workload, relative to other courses, is: 1. Very Light 2. Light 3. Medium 4. Heavy 5. Very Heavy
34. Course pace: 1. Too Slow 2. Slow 3. About Right 4. Fast 5. Too Fast
35. Your expected grade in this course: 1. A 2. B 3. C 4. D 5. Fail
36. In comparison to other courses, how easy is it to get good marks in this course? 1. Very Easy 2. Easy 3. Average 4. Difficult 5. Very Difficult
37. Level of interest in the course before the start of the class: 1. Very Low 2. Low 3. Medium 4. High 5. Very High
BACKGROUND INFORMATION ABOUT THE LECTURER
38. Lecturer's gender: 1. Male 2. Female
39. Lecturer's nationality: 1. Omani 2. Non-Omani Arab 3. Asian (Indian, Pakistani, Philippine, etc) 4. European 5. North American (US American, Canadian) 6. African 7. Other (Please specify)……
40. What is your lecturer's mother tongue: 1. Arabic 2. English 3. An Asian language from the Indian sub-continent or South-west Asia 4. Other (Please specify) ……
OPEN-ENDED COMMENTS
41. Please indicate the important characteristics of this lecturer that have been most valuable to your overall learning experience (particularly aspects not covered by the rating items).
1…… 2…… 3……
42. Please indicate the characteristics of this lecturer that you feel are most important for him/her to improve (particularly aspects not covered by the rating items).
1…… 2…… 3……

Please use the additional space below to clarify any of your responses or to make other comments
…… …… …… …… …… …… …… …… …… …… …… ……

Thank you for completing this questionnaire

APPENDIX 10

Pearson Chi-Square Test

Effect of Student's Gender on Student's Ranking of the Importance of the Characteristics of Effective College Teaching (Significance Level (2-sided): $p<.001$)

Characteristics of Effective Teaching	Chi-Square	*df*	Asymp. Sig.
Encouraging students to participate in classroom activities and discussions	5.693	4	.223
Inviting students to share their ideas and knowledge	5.928	4	.205
Demonstrating good skills in classroom management	5.547	4	.236
Demonstrating very good use of student-centred approaches	1.823	4	.768
Encouraging students to ask questions and ensuring that answers given to students are meaningful	7.831	4	.098
Encouraging students to express their own ideas and/or question the lecturer	13.178	4	.010
Presenting the background or origin of ideas/concepts developed in class	10.207	4	.037
Demonstrating a high level of expressiveness and giving clear explanations	18.491	4	**.001**
Giving lectures/tutorials in a style and pace that facilitate note-taking	38.204	4	**.000**
Showing flexibility and diversity in teaching style	10.021	4	.040
Preparing good course materials and carefully explaining them to students	33.125	4	**.000**
Presenting points of view other than lecturer's own when appropriate	12.159	4	.016
Using lively presentation styles which hold students' interest during class	10.078	4	.039
Making proper use of instructional media, teaching aids, and multi-media labs	9.633	4	.047
Enhancing presentation with the use of humour	2.414	4	.660
Showing strong enthusiasm for the subject	14.094	4	.007
Making the course intellectually challenging and stimulating	14.837	4	.005
Being able to stimulate the interest of the students in the subject	17.681	4	**.001**
Showing dedication to teaching	7.351	4	.118
Being dynamic and energetic in conducting the class	31.014	4	**.000**
Using appropriate and fair methods of evaluating student work	20.353	4	**.000**
Giving valuable feedback on assessments/graded material	5.145	4	.273
Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject	5.958	4	.202
Demonstrating full compliance with the announced objectives of the course	23.344	4	**.000**
Giving assignments/graded materials which test class content	1.870	4	.760

as emphasised by the lecturer			
Having relevant Academic qualifications in teaching English as a second/foreign language	15.098	4	.005
Having sufficient formal teacher training in teaching English as a second/foreign language	17.465	4	.002
Keeping abreast of the latest developments in the field or subject	6.869	4	.143
Having relevant and sufficient experience in teaching English as a second/foreign language	13.689	4	.008
Having full command of the subject matter	28.694	4	**.000**
Making students feel welcome in seeking help/advice in or outside of class	11.919	4	.018
Having a genuine interest in individual students	1.876	4	.759
Showing respect for all students	15.074	4	.005
Being available to students for advice and support in or after class	4.860	4	.302
Showing sensitivity to the culture of the organisation and society at large	8.535	4	.074
Being friendly toward individual students	5.587	4	.232
Having native-like intonation and stress	2.001	4	.736
Being a native speaker of the target language	20.946	4	**.000**

APPENDIX 11

Kruskal-Wallis Test

Effect of Teacher's Ethnicity on Teacher's Ranking of the Importance of the Characteristics of Effective College Teaching (Significance Level: $p<.001$)

Characteristics of Effective Teaching	Chi-Square	*df*	Asymp. Sig.
Encouraging students to participate in classroom activities and discussions	7.035	7	0.425
Inviting students to share their ideas and knowledge	8.812	7	0.266
Demonstrating good skills in classroom management	5.117	7	0.646
Demonstrating very good use of student-centred approaches	14.308	7	0.046
Encouraging students to ask questions and ensuring that answers given to students are meaningful	6.208	7	0.516
Encouraging students to express their own ideas and/or question the lecturer	7.871	7	0.344
Presenting the background or origin of ideas/concepts developed in class	7.289	7	0.399
Demonstrating a high level of expressiveness and giving clear explanations	9.754	7	0.203
Giving lectures/tutorials in a style and pace that facilitate note-taking	8.570	7	0.285
Showing flexibility and diversity in teaching style	9.818	7	0.199
Preparing good course materials and carefully explaining them to students	18.613	7	0.009
Presenting points of view other than lecturer's own when appropriate	5.937	7	0.547
Using lively presentation styles which hold students' interest during class	1.759	7	0.972
Making proper use of instructional media, teaching aids, and multi-media labs	16.216	7	0.023
Enhancing presentation with the use of humour	4.498	7	0.721
Showing strong enthusiasm for the subject	12.733	7	0.079
Making the course intellectually challenging and stimulating	9.426	7	0.224
Being able to stimulate the interest of the students in the subject	5.030	7	0.656
Showing dedication to teaching	25.691	7	**0.001**
Being dynamic and energetic in conducting the class	11.144	7	0.132
Using appropriate and fair methods of evaluating student work	6.988	7	0.430
Giving valuable feedback on assessments/graded material	2.290	7	0.942
Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject	7.575	7	0.372
Demonstrating full compliance with the announced objectives	17.161	7	0.016

of the course			
Giving assignments/graded materials which test class content as emphasised by the lecturer	8.327	7	0.305
Having relevant Academic qualifications in teaching English as a second/foreign language	17.200	7	0.016
Having sufficient formal teacher training in teaching English as a second/foreign language	14.920	7	0.037
Keeping abreast of the latest developments in the field or subject	27.088	7	**0.000**
Having relevant and sufficient experience in teaching English as a second/foreign language	9.632	7	0.210
Having full command of the subject matter	20.400	7	0.005
Making students feel welcome in seeking help/advice in or outside of class	4.539	7	0.716
Having a genuine interest in individual students	15.108	7	0.035
Showing respect for all students	9.355	7	0.228
Being available to students for advice and support in or after class	2.277	7	0.943
Showing sensitivity to the culture of the organisation and society at large	7.497	7	0.379
Being friendly toward individual students	7.155	7	0.413
Having native-like intonation and stress	22.401	7	0.002
Being a native speaker of the target language	36.801	7	**0.000**

APPENDIX 12

Mann-Whitney U Test

Post-hoc Tests for Differences between Rankings of Different Teacher Ethnic Groups of the Importance of *dedication to teaching* (Significance Level(2-tailed): $p<.002$)

Lecturer's Ethnicity	N	Median	Mean Rank	u	z	sig.	r
Omani Arab	21	5	27.86	276.000	-0.854	0.393	.12
Non-Omani Arab	30	5	24.70				
Omani Arab	21	5	51.29	846.000	-1.551	0.121	.14
South Asian or Southwest Asian	96	5	60.69				
Omani Arab	21	5	23.64	133.500	-1.985	0.047	.31
European	19	4	17.03				
Omani Arab	21	5	17.14	129.000	-1.146	0.252	.19
African	15	5	20.40				
Omani Arab	21	5	34.57	366.000	-1.096	0.273	.14
North American	41	4	29.93				
Omani Arab	21	5	18.95	167.000	-0.408	0.683	.07
Southeast Asian	17	5	20.18				
Omani Arab	21	5	14.74	47.500	-1.040	0.298	.20
Other	6	4.5	11.42				
Non-Omani Arab	30	5	50.75	1057.500	-2.829	0.005	.25
South Asian or Southwest Asian	96	5	67.48				
Non-Omani Arab	30	5	26.87	229.000	-1.256	0.209	.18
European	19	4	22.05				
Non-Omani Arab	30	5	20.85	160.500	-1.819	0.069	.27
African	15	5	27.30				
Non-Omani Arab	30	5	36.35	604.500	-0.137	0.891	.02
North American	41	4	35.74				
Non-Omani Arab	30	5	22.50	210.000	-1.146	0.252	.17
Southeast Asian	17	5	26.65				
Non-Omani Arab	30	5	18.88	78.500	-0.541	0.589	.09
Other	6	4.5	16.58				

South Asian or Southwest Asian	96	5	62.27	502.500	-3.937	**0.000**	.37
European	19	4	36.45				
South Asian or Southwest Asian	96	5	55.89	709.500	-0.129	0.898	.01
African	15	5	56.70				
South Asian or Southwest Asian	96	5	75.13	1380.000	-3.461	**0.001**	.30
North American	41	4	54.66				
South Asian or Southwest Asian	96	5	57.76	743.500	-0.806	0.420	.08
Southeast Asian	17	5	52.74				
South Asian or Southwest Asian	96	5	52.54	188.500	-1.944	0.052	.19
Other	6	4.5	34.92				
European	19	4	13.95	75.000	-2.631	0.009	.45
African	15	5	22.00				
European	19	4	26.53	314.000	-1.327	0.185	.17
North American	41	4	32.34				
European	19	4	15.39	102.500	-2.071	0.038	.35
Southeast Asian	17	5	21.97				
European	19	4	12.87	54.500	-0.169	0.866	.03
Other	6	4.5	13.42				
African	15	5	35.10	208.500	-2.111	0.035	.28
North American	41	4	26.09				
African	15	5	17.40	114.000	-0.676	0.499	.12
Southeast Asian	17	5	15.71				
African	15	5	12.10	28.500	-1.619	0.105	.35
Other	6	4.5	8.25				
North American	41	4	27.77	277.500	-1.382	0.167	.18
Southeast Asian	17	5	33.68				
North American	41	4	24.38	107.500	-0.550	0.582	.08
Other	6	4.5	21.42				
Southeast Asian	17	5	12.82	37.000	-1.162	0.245	.24
Other	6	4.5	9.67				

APPENDIX 13

Mann-Whitney U Test

Post-hoc Tests for Differences in Rankings of the Importance of *Keeping Abreast of the latest developments in the field* between the Different Teacher Ethnic Groups (Significance Level (2-tailed): $p<.002$)

Lecturer's Ethnicity	N	Median	Mean Rank	u	z	sig.	r
Omani Arab	21	5	28.26	267.500	-0.972	0.331	.14
Non-Omani Arab	30	4	24.42				
Omani Arab	21	5	65.64	847.500	-1.166	0.244	.11
South Asian or Southwest Asian	95	4	56.92				
Omani Arab	21	5	23.95	127.000	-2.063	0.039	.33
European	19	4	16.68				
Omani Arab	21	5	18.62	155.000	-0.087	0.931	.01
African	15	4	18.33				
Omani Arab	21	5	39.81	235.000	-2.939	0.003	.38
North American	40	4	26.38				
Omani Arab	21	5	19.48	178.000	-0.016	0.987	.00
Southeast Asian	17	4	19.53				
Omani Arab	21	5	15.62	29.000	-2.095	0.036	.40
Other	6	3.5	8.33				
Non-Omani Arab	30	4	62.95	1423.500	-0.010	0.992	.00
South Asian or Southwest Asian	95	4	63.02				
Non-Omani Arab	30	4	27.40	213.000	-1.592	0.111	.23
European	19	4	21.21				
Non-Omani Arab	30	4	21.95	193.500	-0.817	0.414	.12
African	15	4	25.10				
Non-Omani Arab	30	4	42.53	389.000	-2.678	0.007	.32
North American	40	4	30.23				
Non-Omani Arab	30	4	22.38	206.500	-1.182	0.237	.17
Southeast Asian	17	4	26.85				
Non-Omani Arab	30	4	19.90	48.000	-1.931	0.053	.32
Other	6	3.5	11.50				

South Asian or Southwest Asian	95	4	59.97	667.500	-1.954	0.051	.18
European	19	4	45.13				
South Asian or Southwest Asian	95	4	54.45	612.500	-0.951	0.342	.09
African	15	4	62.17				
South Asian or Southwest Asian	95	4	75.28	1208.000	-3.619	**0.000**	.31
North American	40	4	50.70				
South Asian or Southwest Asian	95	4	54.83	648.500	-1.423	0.155	.13
Southeast Asian	17	4	65.85				
South Asian or Southwest Asian	95	4	52.45	147.000	-2.178	0.029	.22
Other	6	3.5	28.00				
European	19	4	14.74	90.000	-1.923	0.054	.33
African	15	4	21.00				
European	19	4	31.79	346.000	-0.587	0.557	.08
North American	40	4	29.15				
European	19	4	14.84	92.000	-2.387	0.017	.40
Southeast Asian	17	4	22.59				
European	19	4	13.55	46.500	-0.715	0.475	.14
Other	6	3.5	11.25				
African	15	4	37.13	163.000	-2.731	0.006	.37
North American	40	4	24.58				
African	15	4	16.23	123.500	-0.166	0.868	.03
Southeast Asian	17	4	16.74				
African	15	4	12.70	19.500	-2.097	0.036	.46
Other	6	3.5	6.75				
North American	40	4	24.48	159.000	-3.376	**0.001**	.45
Southeast Asian	17	4	39.65				
North American	40	4	23.78	109.000	-0.384	0.701	.06
Other	6	3.5	21.67				
Southeast Asian	17	4	13.97	17.500	-2.576	0.010	.54
Other	6	3.5	6.42				

APPENDIX 14

Mann-Whitney U Test

Post-hoc Tests for Differences between Rankings of Different Teacher Ethnic Groups of the Importance of *being a native speaker of the target language*
(Significance Level (2-tailed): *p*<.002)

Lecturer's Ethnicity	N	Median	Mean Rank	u	z	sig.	r
Omani Arab	21	3	26.60	302.500	-0.247	0.805	.03
Non-Omani Arab	30	2.5	25.58				
Omani Arab	21	3	69.57	765.000	-1.776	0.076	.16
South Asian or Southwest Asian	95	2	56.05				
Omani Arab	21	3	18.81	164.000	-0.989	0.323	.16
European	19	3	22.37				
Omani Arab	21	3	18.38	155.000	-0.083	0.934	.01
African	15	3	18.67				
Omani Arab	21	3	25.38	302.000	-1.849	0.064	.24
North American	40	3	33.95				
Omani Arab	21	3	20.43	159.000	-0.597	0.550	.10
Southeast Asian	17	3	18.35				
Omani Arab	21	3	13.86	60.000	-0.181	0.856	.03
Other	6	2	14.50				
Non-Omani Arab	30	2.5	75.83	1040.000	-2.350	0.019	.21
South Asian or Southwest Asian	95	2	58.95				
Non-Omani Arab	30	2.5	22.30	204.000	-1.720	0.085	.25
European	19	3	29.26				
Non-Omani Arab	30	2.5	22.30	204.000	-0.524	0.600	.08
African	15	3	24.40				
Non-Omani Arab	30	2.5	27.27	353.000	-3.032	**0.002**	.36
North American	40	3	41.68				
Non-Omani Arab	30	2.5	24.57	238.000	-0.393	0.695	.06
Southeast Asian	17	3	23.00				
Non-Omani Arab	30	2.5	18.53	89.000	-0.044	0.965	.01
Other	6	2	18.33				

South Asian or Southwest Asian	95	2	52.78	454.000	-3.591	**0.000**	-0.34
European	19	3	81.11				
South Asian or Southwest Asian	95	2	53.12	486.000	-2.098	0.036	.20
African	15	3	70.60				
South Asian or Southwest Asian	95	2	56.76	832.500	-5.371	**0.000**	.46
North American	40	3	94.69				
South Asian or Southwest Asian	95	2	54.91	656.000	-1.313	0.189	.12
Southeast Asian	17	3	65.41				
South Asian or Southwest Asian	95	2	50.39	227.000	-0.892	0.373	.09
Other	6	2	60.67				
European	19	3	18.84	117.000	-0.917	0.359	.16
African	15	3	15.80				
European	19	3	26.68	317.000	-1.061	0.289	.14
North American	40	3	31.58				
European	19	3	21.37	107.000	-1.794	0.073	.30
Southeast Asian	17	3	15.29				
European	19	3	13.53	47.000	-0.651	0.515	.13
Other	6	2	11.33				
African	15	3	21.70	205.500	-1.855	0.064	.25
North American	40	3	30.36				
African	15	3	17.83	107.500	-0.792	0.428	.14
Southeast Asian	17	3	15.32				
African	15	3	11.07	44.000	-0.080	0.936	.02
Other	6	2	10.83				
North American	40	3	32.84	186.500	-2.790	0.005	.37
Southeast Asian	17	3	19.97				
North American	40	3	24.23	91.000	-0.976	0.329	.14
Other	6	2	18.67				
Southeast Asian	17	3	11.82	48.000	-0.218	0.827	.05
Other	6	2	12.50				

APPENDIX 15

Kruskal-Wallis Test

Effect of Teacher's Mother Tongue on Teacher's Ranking of the Importance of the Characteristics of Effective College Teaching (Significance Level: $p<.001$)

Characteristics of Effective Teaching	Chi-Square	*df*	Asymp. Sig.
Encouraging students to participate in classroom activities and discussions	12.266	10	0.268
Inviting students to share their ideas and knowledge	12.745	10	0.238
Demonstrating good skills in classroom management	19.030	10	0.040
Demonstrating very good use of student-centred approaches	18.597	10	0.046
Encouraging students to ask questions and ensuring that answers given to students are meaningful	6.593	10	0.763
Encouraging students to express their own ideas and/or question the lecturer	15.419	10	0.118
Presenting the background or origin of ideas/concepts developed in class	19.498	10	0.034
Demonstrating a high level of expressiveness and giving clear explanations	10.626	10	0.387
Giving lectures/tutorials in a style and pace that facilitate note-taking	4.241	10	0.936
Showing flexibility and diversity in teaching style	12.469	10	0.255
Preparing good course materials and carefully explaining them to students	17.818	10	0.058
Presenting points of view other than lecturer's own when appropriate	7.823	10	0.646
Using lively presentation styles which hold students' interest during class	4.785	10	0.905
Making proper use of instructional media, teaching aids, and multi-media labs	20.954	10	0.021
Enhancing presentation with the use of humour	5.012	10	0.890
Showing strong enthusiasm for the subject	14.997	10	0.132
Making the course intellectually challenging and stimulating	14.474	10	0.152
Being able to stimulate the interest of the students in the subject	4.237	10	0.936
Showing dedication to teaching	21.732	10	0.017
Being dynamic and energetic in conducting the class	9.345	10	0.500
Using appropriate and fair methods of evaluating student work	4.463	10	0.924
Giving valuable feedback on assessments/graded material	9.611	10	0.475
Assigning homework/readings which are valuable and contribute to appreciation and understanding of the subject	9.953	10	0.445
Demonstrating full compliance with the announced objectives of the course	8.307	10	0.599
Giving assignments/graded materials which test class content	7.529	10	0.675

as emphasised by the lecturer			
Having relevant Academic qualifications in teaching English as a second/foreign language	37.766	10	**0.000**
Having sufficient formal teacher training in teaching English as a second/foreign language	18.451	10	0.048
Keeping abreast of the latest developments in the field or subject	28.271	10	0.002
Having relevant and sufficient experience in teaching English as a second/foreign language	10.622	10	0.388
Having full command of the subject matter	11.709	10	0.305
Making students feel welcome in seeking help/advice in or outside of class	6.719	10	0.752
Having a genuine interest in individual students	13.754	10	0.185
Showing respect for all students	13.342	10	0.205
Being available to students for advice and support in or after class	5.422	10	0.861
Showing sensitivity to the culture of the organisation and society at large	5.578	10	0.849
Being friendly toward individual students	14.455	10	0.153
Having native-like intonation and stress	22.409	10	0.013
Being a native speaker of the target language	55.583	10	**0.000**

APPENDIX 16

Mann-Whitney U Test

Post-hoc Tests for Differences between the Different Teacher L1 Groups in Their Perceptions of the Importance of *having relevant academic qualifications* (Significance Level (2-tailed): p<.001)

L1	N	Median	Mean Ranks	U	Z	Sig.	r
Arabic	49	5	77.77	1089.500	-4.052	**0.000**	.36
English	75	4	52.53				
Arabic	49	5	35.21	185.500	-2.869	0.004	.36
Urdu	14	4	20.75				
Arabic	49	5	37.10	289.000	-2.375	0.018	.29
Hindi	18	4	25.56				
Arabic	49	5	35.65	522.000	-1.531	0.126	.18
Malayalam	26	5	42.42				
Arabic	49	5	35.79	353.500	-1.390	0.164	.17
Tamil	18	4	29.14				
Arabic	49	5	35.74	306.500	-1.809	0.070	.22
Tagalog	17	4	27.03				
Arabic	49	5	32.26	330.500	-0.237	0.813	.03
Other South Asian or Southwest Asian language	14	5	31.11				
Arabic	49	5	28.06	95.000	-0.933	0.351	.13
An African language	5	4	22.00				
Arabic	49	5	27.10	93.000	-0.193	0.847	.03
A European language	4	5	25.75				
Arabic	49	5	25.59	29.000	-1.125	0.261	.16
Other	2	5	36.00				
English	75	4	44.81	510.500	-0.170	0.865	.02
Urdu	14	4	46.04				
English	75	4	46.19	614.500	-0.613	0.540	.06
Hindi	18	4	50.36				
English	75	4	44.03	452.500	-4.293	**0.000**	.43
Malayalam	26	5	71.10				

English	75	4	44.56	492.000	-1.860	0.063	.19
Tamil	18	4	57.17				
English	75	4	44.41	481.000	-1.647	0.100	.17
Tagalog	17	4	55.71				
English	75	4	42.76	357.000	-1.981	0.048	.21
Other South Asian or Southwest Asian language	14	5	57.00				
English	75	4	39.93	145.000	-0.880	0.379	.10
An African language	5	4	49.00				
English	75	4	39.20	90.000	-1.398	0.162	.16
A European language	4	5	55.00				
English	75	4	38.31	23.000	-1.735	0.083	.20
Other	2	5	65.00				
Urdu	14	4	15.82	116.500	-0.381	0.703	.07
Hindi	18	4	17.03				
Urdu	14	4	12.61	71.500	-3.534	**0.000**	.56
Malayalam	26	5	24.75				
Urdu	14	4	13.75	87.500	-1.580	0.114	.28
Tamil	18	4	18.64				
Urdu	14	4	13.57	85.000	-1.498	0.134	.27
Tagalog	17	4	18.00				
Urdu	14	4	12.18	65.500	-1.598	0.110	.30
Other South Asian or Southwest Asian language	14	5	16.82				
Urdu	14	4	9.39	26.500	-0.840	0.401	.19
An African language	5	4	11.70				
Urdu	14	4	8.57	15.000	-1.478	0.139	.35
A European language	4	5	12.75				
Urdu	14	4	7.71	3.000	-1.852	0.064	.46
Other	2	5	14.00				
Hindi	18	4	16.14	119.500	-3.103	0.002	.47
Malayalam	26	5	26.90				
Hindi	18	4	16.78	131.000	-1.049	0.294	.17

Tamil	18	4	20.22				
Hindi	18	4	16.64	128.500	-0.877	0.380	.15
Tagalog	17	4	19.44				
Hindi	18	4	14.72	94.000	-1.307	0.191	.23
Other South Asian or Southwest Asian language	14	5	18.79				
Hindi	18	4	11.64	38.500	-0.511	0.610	.11
An African language	5	4	13.30				
Hindi	18	4	10.83	24.000	-1.080	0.280	.23
A European language	4	5	14.50				
Hindi	18	4	9.83	6.000	-1.598	0.110	.36
Other	2	5	16.50				
Malayalam	26	5	25.94	144.500	-2.489	0.013	.38
Tamil	18	4	17.53				
Malayalam	26	5	25.92	119.000	-2.921	0.003	.45
Tagalog	17	4	16.00				
Malayalam	26	5	21.75	149.500	-1.176	0.240	.19
Other South Asian or Southwest Asian language	14	5	18.18				
Malayalam	26	5	16.98	39.500	-1.724	0.085	.31
An African language	5	4	10.90				
Malayalam	26	5	16.00	39.000	-1.029	0.303	.19
A European language	4	5	12.25				
Malayalam	26	5	14.27	20.000	-0.750	0.454	.14
Other	2	5	17.50				
Tamil	18	4	18.44	145.000	-0.300	0.764	.05
Tagalog	17	4	17.53				
Tamil	18	4	15.67	111.000	-0.622	0.534	.11
Other South Asian or Southwest Asian language	14	5	17.57				
Tamil	18	4	12.11	43.000	-0.164	0.870	.03
An African language	5	4	11.60				
Tamil	18	4	11.17	30.000	-0.568	0.570	.12

A European language	4	5	13.00				
Tamil	18	4	9.89	7.000	-1.532	0.126	.34
Other	2	5	16.00				
Tagalog	17	4	14.82	99.000	-0.863	0.388	.15
Other South Asian or Southwest Asian language	14	5	17.43				
Tagalog	17		11.50	42.500	0.000	1.000	.00
An African language	5	4	11.50				
Tagalog	17	4	10.53	26.000	-0.840	0.401	.18
A European language	4	5	13.00				
Tagalog	17	4	9.29	5.000	-1.831	0.067	.42
Other	2	5	16.00				
Other South Asian or Southwest Asian language	14	5	10.36	30.000	-0.522	0.602	.12
An African language	5	4	9.00				
Other South Asian or Southwest Asian language	14	5	9.50	28.000	0.000	1.000	.00
A European language	4	5	9.50				
Other South Asian or Southwest Asian language	14	5	8.14	9.000	-0.976	0.329	.24
Other	2	5	11.00				
An African language	5	4	4.60	8.000	-0.537	0.592	.18
A European language	4	5	5.50				
An African language	5	4	3.40	2.000	-1.296	0.195	.49
Other	2	5	5.50				
A European language	4	5	3.00	2.000	-1.118	0.264	.46
Other	2	5	4.50				

APPENDIX 17

Mann-Whitney U Test

Post-hoc Tests for Differences between the Different Teacher L1 Groups in Their Perceptions of the Importance of *Being a Native Speaker of English* (Significance Level (2-tailed): p<.001)

L1	N	Median	Mean Ranks	U	Z	Sig.	r
Arabic	49	3	50.91	1269.500	-2.888	0.004	.26
English	74	3	69.34				
Arabic	49	3	33.21	332.500	-0.574	0.566	.07
Urdu	15	3	30.17				
Arabic	49	3	34.84	351.000	-0.992	0.321	.12
Hindi	17	2	29.65				
Arabic	49	3	44.94	297.000	-3.973	**0.000**	.46
Malayalam	26	1	24.92				
Arabic	49	3	37.38	275.500	-2.433	0.015	.30
Tamil	18	1	24.81				
Arabic	49	3	34.47	369.000	-0.721	0.471	.09
Tagalog	17	3	30.71				
Arabic	49	3	34.40	225.500	-2.007	0.045	.25
Other South Asian or Southwest Asian language	14	1.5	23.61				
Arabic	49	3	28.38	79.500	-1.327	0.185	.18
An African language	5	1	18.90				
Arabic	49	3	27.51	73.000	-0.870	0.384	.12
A European language	4	2	20.75				
Arabic	49	3	26.01	48.500	-0.025	0.980	.00
Other	2	2.5	25.75				
English	74	3	47.99	333.500	-2.506	0.012	.27
Urdu	15	3	30.23				
English	74	3	49.76	351.000	-2.912	0.004	.31
Hindi	17	2	29.65				
English	74	3	60.28	238.500	-5.839	**0.000**	.58
Malayalam	26	1	22.67				

English	74	3	51.93	264.500	-4.061	**0.000**	.42
Tamil	18	1	24.19				
English	74	3	49.53	368.000	-2.744	0.006	.29
Tagalog	17	3	30.65				
English	74	3	48.61	213.500	-3.564	**0.000**	.38
Other South Asian or Southwest Asian language	14	1.5	22.75				
English	74	3	41.51	73.500	-2.316	0.021	.26
An African language	5	1	17.70				
English	74	3	40.57	69.000	-1.848	0.065	.21
A European language	4	2	19.75				
English	74	3	38.74	56.500	-0.586	0.558	.07
Other	2	2.5	29.75				
Urdu	15	3	17.13	118.000	-0.376	0.707	.07
Hindi	17	2	15.94				
Urdu	15	3	27.00	105.000	-2.710	0.007	.42
Malayalam	26	1	17.54				
Urdu	15	3	19.70	94.500	-1.582	0.114	.28
Tamil	18	1	14.75				
Urdu	15	3	16.67	125.000	-0.100	0.920	.02
Tagalog	17	3	16.35				
Urdu	15	3	16.90	76.500	-1.311	0.190	.24
Other South Asian or Southwest Asian language	14	1.5	12.96				
Urdu	15	3	11.17	27.500	-0.935	0.350	.21
An African language	5	1	8.50				
Urdu	15	3	10.33	25.000	-0.534	0.593	.12
A European language	4	2	8.75				
Urdu	15	3	8.90	13.500	-0.235	0.814	.06
Other	2	2.5	9.75				
Hindi	17	2	27.26	131.500	-2.469	0.014	.38
Malayalam	26	1	18.56				
Hindi	17	2	20.21	115.500	-1.332	0.183	.23

Tamil	18	1	15.92				
Hindi	17	2	17.03	136.500	-0.289	0.772	.05
Tagalog	17	3	17.97				
Hindi	17	2	17.44	94.500	-1.026	0.305	.18
Other South Asian or Southwest Asian language	14	1.5	14.25				
Hindi	17	2	12.00	34.000	-0.704	0.481	.15
An African language	5	1	9.80				
Hindi	17	2	11.18	31.000	-0.283	0.777	.06
A European language	4	2	10.25				
Hindi	17	2	9.91	15.500	-0.207	0.836	.05
Other	2	2.5	10.75				
Malayalam	26	1	21.35	204.000	-0.852	0.394	.13
Tamil	18	1	24.17				
Malayalam	26	1	18.25	123.500	-2.692	0.007	.41
Tagalog	17	3	27.74				
Malayalam	26	1	19.04	144.000	-1.250	0.211	.20
Other South Asian or Southwest Asian language	14	1.5	23.21				
Malayalam	26	1	15.58	54.000	-0.716	0.474	.13
An African language	5	1	18.20				
Malayalam	26	1	14.92	37.000	-1.098	0.272	.20
A European language	4	2	19.25				
Malayalam	26	1	14.15	17.000	-0.973	0.331	.18
Other	2	2.5	19.00				
Tamil	18	1	15.61	110.000	-1.536	0.124	.26
Tagalog	17	3	20.53				
Tamil	18	1	16.06	118.000	-0.338	0.736	.06
Other South Asian or Southwest Asian language	14	1.5	17.07				
Tamil	18	1	11.92	43.500	-0.128	0.898	.03
An African language	5	1	12.30				
Tamil	18	1	11.22	31.000	-0.484	0.629	.10

A European language	4	2	12.75				
Tamil	18	1	10.25	13.500	-0.643	0.520	.14
Other	2	2.5	12.75				
Tagalog	17	3	17.76	89.000	-1.259	0.208	.23
Other South Asian or Southwest Asian language	14	1.5	13.86				
Tagalog	17	3	12.12	32.000	-0.885	0.376	.19
An African language	5	1	9.40				
Tagalog	17	3	11.29	29.000	-0.481	0.630	.10
A European language	4	2	9.75				
Tagalog	17	3	9.88	15.000	-0.280	0.779	.06
Other	2	2.5	11.00				
Other South Asian or Southwest Asian language	14	1.5	10.04	34.500	-0.051	0.960	.01
An African language	5	1	9.90				
Other South Asian or Southwest Asian language	14	1.5	9.29	25.000	-0.344	0.731	.08
A European language	4	2	10.25				
Other South Asian or Southwest Asian language	14	1.5	8.32	11.500	-0.428	0.669	.11
Other	2	2.5	9.75				
An African language	5	1	4.80	9.000	-0.283	0.777	.09
A European language	4	2	5.25				
An African language	5	1	3.70	3.500	-0.648	0.517	.24
Other	2	2.5	4.75				
A European language	4	2	3.25	3.000	-0.500	0.617	.20
Other	2	2.5	4.00				

REFERENCES

Abrami, P.C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education*, *30*: 221-270.

Abrami, P.C. & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, & J.Franklin (Eds.), *Student ratings of instruction: Issues for improving practice: New directions for teaching and learning* (pp.97-111). San Francisco: Jossey-Bass.

Abrami, P.C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness- Generalizability of "N=1" research: comment on Marsh (1991). *Journal of Educational Psychology*, *83*: 411-415.

Abrami, P. C., and d'Apollonia, S. (1998). *The positive relationship between course grades and course ratings: What is the cause and what, if anything, can be done about it?*. Debate presented at the 79th Annual Meeting of the American Educational Research Association, San Diego, California, April 1998.

Abrami, P.C., d'Apollonia, S., & Cohen, P.A. (1990) Validity of student ratings of instruction: What we know and what we do not know. *Journal of Educational Psychology*, *82* (2): 219-231.

Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. E Perry & J.C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321-367). New York: Agathon Press.

Abrami, P. C., Perry, R. P., & Leventhal, L. (1982). The relationship between student personality characteristics, teacher ratings, and student achievement. *Journal of Educational Psychology, 74*: 111-125.

Adams, J. V. (1997). Student evaluations: The ratings game. *Inquiry*, *1*(2): 10-16.

Ahmadi, M., Helms, M. M., & Raiszadeh, F. (2001). Business students' perceptions of faculty evaluations. *The International Journal of Educational Management*, *15*(1): 12-22.

Al Musawi, A. S. & Abdelraheem, A. Y. (2004). E-learning at Sultan Qaboos University: status and future. *British Journal of Educational Technology*, *35*(3): 363–367.

Al Shmeli, S. (2009). *Higher education in the Sultanate of Oman: Planning in the context of globalisation.* Paper presented at the IIEP Policy Forum: Tertiary Education in Small States: Planning in the Context of Globalization. UNESCO, International Institute for Educational Planning, 2-3 July 2009.

Al-Arishi, A.Y. (1994). Role-play, real-play, and surreal-play in the ESOL classroom. *LT Journal, 48* (4): 337-346.

Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of Teacher Evaluation* (pp. 110-145). Beverly Hills, CA: Sage.

Aleamoni, L. M. (1987). Student ratings myths versus research facts. J*ournal of Personnel Evaluation in Education, 1*: 111-119.

Aleamoni, L. M. (1999) Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2): 153-166.

Aleamoni, L.M. & Hexner, P.Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science, 9*: 67-84.

Aleamoni, L.M., & Thomas, G.S. (1980). Differential relationships of student, instructor, and course characteristics to general and specific items on a course evaluation questionnaire. *Teaching of Psychology, 7*(4): 233-235.

Al-Hinai, N. (2005). *Foundation English language programmes at the Colleges of Technology, Ministry of Manpower*. Paper presented at Higher Education Workshop on Quality Assurance in Foundation Year Programmes: Realities and Challenges. Golden Tulip Hotel, Khasab, November 13-15, 2005.

Al-Husseini, S. (2006). The visible and invisible role of English foundation programmes: A search for communication opportunities within EFL contexts. *Asian EFL Journal, 8*(4). Retrieved 22 January, 2008 from http://www.asian-efl-journal.com

Al-Issa, A. (2005). An Ideological discussion of the impact of the NNESTs' English language knowledge on Omani ESL policy implementation: A special reference to the Omani context. *Asian EFL Journal, 7*(3). Retrieved 22 January, 2008 from http://www.asian-efl-journal.com

Al-Issa, A., & Sulieman, H. (2007). Student evaluations of teaching: perceptions and biasing factors. *Quality Assurance in Education*, *15*(3): 302-317.

Alweshahi, Y., Harley, D., and Cook, D. A. (2007). Students' perception of the characteristics of effective bedside teachers. *Medical Teacher*, *29* (2): 204-209.

Amin, N. (2000). Negotiating nativism: Minority immigrant women ESL teachers and the native speaker construct. Unpublished doctoral dissertation, University of Toronto, Canada.

Anderson, K., & Miller, E.D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics, 30*(2): 216-219.

Anderson, S. E. & Gansneder, B. M. (1995). Using electronic mail surveys and computer-monitored data for studying computer-mediated communication systems. *Social Science Computer Review, 13* (1): 33-46.

Armstrong, J. S. (1998). Are student ratings of instruction useful?, *American Psychologist*, *53*(11): 1223-1224.

Arreola, R. A. (1986). Evaluating the dimensions of teaching. *Instructional Evaluation*, *8*: 4-12.

Arreola, R. A. (1989). Defining and evaluating the elements of teaching. *Proceedings of Academic Chairpersons: Evaluating Faculty, Students, and Programs* (pp.1-14). Manhattan: Kansas State University.

Axinn, W.G., Fricke, T., and Thornton, A. (1991). The microdemographic community study approach: Improving data quality by integrating the ethnographic method. *Sociological Methods and Research, 20*(2): 187-217.

Ayidiya, S. A. & Mckee, J. M. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*(2): 229-247.

Baba, V. V., & Ace, M. E. (1989). Serendipity in leadership, initiating structure and consideration in the classroom, *Human Relations*, *42*: 509-525.

Babbar, S. (1995). Applying total quality management to educational instruction: A case study from a US public university. *International Journal of Public Sector Management*, *8*(7): 35-55.

Bachmann, D., Elfrink, J., & Vazzana, G. (1996). Tracking the progress of e-mail vs. snail mail. *Marketing Research, 8*(2): 30-35.

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business

program in the UAE. *International Journal of Educational Management*, *20* (1): 43-59.

Balam, E. & Shannon, D. (2010). Student ratings of college teaching: A comparison of faculty and their students. *Assessment and Evaluation in Higher Education, 35*(2): 209-221.

Barnes, L. B., & Barnes, M. W. (1993). Academic discipline and generalizability of student evaluations of instruction. *Research in Higher Education*, *34*(2): 135-149.

Basow, S. A., and Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology*, *79*(3): 308-314.

Baum, P., and Brown, W. W. (1980). Student and faculty perceptions of teaching effectiveness. *Research in Higher Education*, *13* (3): 233-242.

Bennett, K. (1994). Going native. *Practical English Teaching, 14* (3): 10-11.

Bernard, H. R. (2000). *Social research methods: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage Publications.

Birbili, M. (2000). Translating from one language to another. *Social Research Update*, *31*(Winter): 1-7. Retrieved 20 July 2007 from http://sru.soc.surrey.ac.uk/SRU31.html

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Princeton, N.J.: Carnegie Foundation for the Advancement of Teaching.

Braine, G. (Ed.) (1999). *Non-native educators in English language teaching*. Mahwah, NJ: Erlbaum.

Braskamp, L.A., & Ory, J.C. (1994) *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco: Jossey-Bass.

Broder, J.M., & Dorfman, J.H. (1994). Determinants of teaching quality: What's important to students? *Research in Higher Education, 35*(2): 235-249.

Brown, J. D. (1998). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Brown, J.D. (2001).*Using surveys in language programs*. Cambridge: Cambridge University Press.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford: Oxford University Press.

Brown, R. (1996). Teaching profiles: The quality context. In R. Aylett, & K. Gregory (Eds.), *Evaluating teacher quality in higher education*. London: The Falmer Press.

Brumfit, C. (1984). *Communicative methodology in language teaching: The roles of fluency and accuracy*. Cambridge: Cambridge University Press.

Bryman, A. (2004). S*ocial research methods (2nd ed.)*. New York: Oxford University Press.

Buck, D. (1998). Student evaluation of teaching measure the intervention, not the effect. *American Psychologist*, *53*(11): 1224-1226.

Canagarajah, S. (1999). Interrogating the native speaker fallacy: Non-linguistic roots, non-pedagogical results. In G. Braine (Ed.), *Non-native educators in English language teaching* (pp. 77-92). Mahwah, NJ: Erlbaum.

Carroll, M., Razvi, S., & Goodliffe, T. (2009). *Using foundation program academic standards as a quality enhancement tool.* A paper presented at the International Network for Quality Assurance Agencies in Higher Education (INQAAHE) Conference. Abu Dhabi, UAE, March 30-April 2, 2009.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research*. IDEA Paper No. 20. Manhattan, KS: Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University. Retrieved July, 25, 2007 from http: //www. idea.ksu.edu.

Cashin, W.E. (1989). *Defining and evaluating college teaching*. IDEA Paper No. 21. Manhattan, KS: Centre for faculty Evaluation and Development, Division of Continuing Education, Kansas State University. Retrieved August, 16, 2007 from http: //www. idea.ksu.edu.

Cashin, W.E. (1995) *Student ratings of teaching: The research revisited*. IDEA Paper No. 32. Manhattan, KS: Centre for faculty Evaluation and Development, Division of Continuing Education, Kansas State University. Retrieved August, 16, 2007 from http: // www.idea.ksu.edu.

Cashin, W. E. (1996). *Developing an effective faculty evaluation system*. IDEA Paper No. 33. Manhattan, KS: Centre for faculty Evaluation and Development, Division of Continuing Education, Kansas State University. Retrieved August, 29, 2007 from http: // www.idea.ksu.edu.

Cashin, W.E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology*, *84*(4): 563-572.

Cashin, W.E., Downey, R.G., & Sixbury, G.R.(1994). Global and specific ratings of teaching effectiveness and their relation to course objectives: Reply to Marsh (1994). *Journal of Educational Psychology, 86*: 649-657.

Centra, J. A. (1975). Colleagues as raters of classroom instruction, *Journal of Higher Education*, *46*: 327-337.

Centra, J.A. (1977). *How universities evaluate faculty performance: A survey of department heads*. *GREB Research Report, No. 75-56R*. Princeton, N.J.: Educational Testing Service.

Centra, J.A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching & Determining faculty Effectiveness*. California: Jossey-Bass Inc.

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5): 495-518.

Centra, J., Froh, R. C., Gray, P. J., and Lambert, L. M. (1987). *A Guide to Evaluating Teaching for Promotion and Tenure*. Acton, Mass.: Copley Publishing Group.

Centra, J.A., & Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*(1): 17-33.

Chen, A., Liu, Z., & Ennis, C. D. (1997). Universality and uniqueness of teacher educational value orientations: A cross-cultural comparison between USA and China. *Journal of Research and Development in Education, 30*: 135–143.

Chowdhury, R. and Le Ha, P. (2008). Reflecting on western TESOL training and communicative language teaching: Bangladeshi teachers' voices. *Asia Pacific Journal of Education, 28* (3): 305–316.

Clarkson, P.C. (1984). Papua New Guinea students' perceptions of mathematics lecturers. *Journal of Educational Psychology*, *76*: 1386-1395.

Clough, P., & Nutbrown, C. (2007). *A student's guide to methodology*. London: Sage Publications Ltd.

Coffey, M. & Gibbs, G. (2001). The evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education*, *26* (1): 89-93.

Cohen, J.W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London: Routledge Falmer.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college teaching: A meta-analysis of findings. *Research in Higher Education, 13*: 321-341.

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies. *Review of Educational Research*, *51*(3): 281-309. Retrieved 31 August, 2007 from http://links.jstor.org.

Coll, R. K., & Chapman, R. (2000). Choices of methodology for cooperative education researchers. *Asia-Pacific Journal of Cooperative Education, 1* (1): 1-8.

Conrad, C. F. (1982). Grounded theory: An alternative approach to research in higher education. *The Review of Higher Education, 5*(4): 239-249.

Coomber, R. (2002). Signing your life away?: Why Research Ethics Committees (REC) shouldn't always require written confirmation that participants in research have been informed of the aims of a study and their rights - the case of criminal populations (Commentary). *Sociological Research Online, 7*(1). Retrieved 14 February, 2008 from http://www.socresonline.org.uk/7/1/coomber.html

Costello, A.B. & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7): 1-9.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41*(5): 511-535. Retrieved 31 August, 2007 from www.jstor.org.

Cranton, P. and Smith, R.A. (1986). A new look at the effect of course characteristics on student ratings of instruction. *American Educational Research Journal*, Spring: 117-28.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approach* (2nd ed.). London: Sage Publications.

Creswell, J. W. (2005). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (2nd ed.). New Jersey: Merrill Prentice Hall.

Creswell, J. W., Fetters, M. D., & Ivankova, N. V. (2004). Designing a mixed methods study in primary care. *Annals of Family Medicine, 2*(1): 7-12.

Creswell, J. W. & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). London: Sage Publications.

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

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*: 297-334.

Crowl, T.K.(1996). *Fundamentals of educational research* (2nd ed.). Boston: McGraw-Hill.

Crumbley, L., & Fliedner, E. (2002). Accounting administrators' perceptions of student evaluation of teaching (SET) information. *Quality Assurance in Education*, *10* (4): 213-222.

Crumbley, L., Henry, B. K., & Kratchman, S. H. (2001). Students' perceptions of the evaluation of college teaching. *Quality Assurance in Educa*tion, *9*(4): 197-207.

Cuthbert, P. F. (1996). Managing service quality in HE: Is SERVQUAL the answer? Part I. *Managing Service Quality*, *6*(2): 11-16.

d'Apollonia, S., & Abrami, P. (1996). *Variables moderating the validity of student ratings of instruction: A meta-analysis*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA, April 1996.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11): 1198-1208.

Daly, D. J. (2000). Student rating teachers is idiocy. *Omaha World- Herald*, 26 March 2000. Retrieved 12 September, 2007 from http://www.bus.lsu.edu/accounting/faculty/lcrumbley/students_rating.htm

Darling-Hammond, L., Wise, A.E., Pease, S.R. (1983) Teacher Evaluation in the Organizational Context: A Review of the Literature. *Review of Educational Research*, *53*(3): 285-328.

Dee, T.S. (2005). A teacher like me: Does race, ethnicity, or gender matter? Papers and Proceedings of the One Hundred Seventeenth Annual Meeting of the American Economic Association, Philadelphia, PA, January 7-9. *The American Economic Review, 95*(2): 158-165.

Denzin, N. K., & Lincoln, Y. (Eds.). (2000). *Handbook of qualitative research*. London: Sage Publications.

Dillman, D. A., Sinclair, M. D., & Clark, J. R. (1993). Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. *The Public Opinion Quarterly, 57* (3): 289-304.

Dilts, D.A., Haber, L.J., Bialik, D. (1994) *Assessing What Professors Do: An Introduction to Academic Performance Appraisal in Higher Education*. London: Greenwood Press.

Dommeyer, C. J. & Moriarty, E. (2000). Comparing two forms of an e-mail survey: embedded vs. attached, International Journal of Market Research, *42*(1): 39-50.

Donaldson, J. F. & Flannery, D. (1993). A triangulated study comparing adult college students perceptions of effective teaching with those of traditional students. *Continuing Higher Education Review*, *57*(3): 147-165.

Dowell, D.A., & Neal, J.A., (1983). The validity and accuracy of student ratings of instruction: A reply to Peter A. Cohen. *Journal of Higher Education*, *54*: 459-463.

Dwinell, P. L., & Higbee, J. J. (1993). Students' perceptions of the value of teaching evaluations. *Perceptual and Motor Skills, 76:* 995-1000.

Ehrenberg, R.G., Goldhaber, D. D., and Brewer, D. J. (1995). Do teachers' race, gender and ethnicity matter? Evidence from the national educational longitudinal study of 1988. *Industrial and Labor Relations Review, 48*(3): 547-561.

Ellett, C. D. & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education*, *17*(1): 101-128.

Elton, L. (1984). Evaluating teaching and assessing teachers in universities. *Assessment and Evaluation in Higher Education*, *9*(2): 97-115.

Elton, L. (1996). Criteria for teaching competence and teaching excellence in higher education. In R. Aylett, & K. Gregory (Eds.), *Evaluating teacher quality in higher education*. London: The Falmer Press.

Emery, C.R., Kramer, T.C., & Tian, R.G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality assurance in Education*, 11(1): 37-46.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*: 543-553.

Farquharson, M. (1989). *Learning styles of Arab students in EFL classrooms.* A paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages, San Antonio, Texas, March 7-11, 1989.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, *5*(3): 243-288.

Feldman, K.A. (1977). Consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 6*: 223-274.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers- what we know and what we don't. *Research in Higher Education*, *9* (3): 199-242.

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, *10*(2): 149-172.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluation they receive from students. *Research in Higher Education*, *18*(1): 3-124.

Feldman, K. A. (1984). Class size and college students' evaluation of teachers and courses: A closer look. *Research in Higher Education*, *21*(4): 45-116.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, *24*(2): 139-213.

Feldman, K.A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education*, *26*: 227-298.

Feldman, K. A. (1988). Effective college teaching from students' and faculty's views: Matched or mismatched priorities. *Research in Higher Education*, *28* (4): 291-344.

Feldman, K,A. (1989a). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multi-section validity studies. *Research in Higher Education*, *30* (6): 583-645.

Feldman, K.A. (1989b). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers. *Research in Higher Education*, *30*: 137-194.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I- Evidence from the social laboratory and experiments. *Research in Higher Education*, *33*(3): 317-375.

Feldman, K. A. (1993). College students' views of male and female college teachers. Part II – Evidence from student evaluations of their classroom teachers. *Research in Higher Education*, *34* (2): 151-211.

Feldman, K.A. (1997). Identifying exemplary teachers and teaching: evidence from student ratings. In R.P. Perry and J.C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp.368-395). New York: Agathon Press.

Feldman, K.A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In: J.C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (pp. 35-74). New York: Agathon Press.

Fernandez, J. & Mateo, M.A. (1997). Student and faculty gender in rating of university teaching quality. *Sex Roles: A Journal of Research, 37* (11-12): 997-1003.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Sage publications: London.

Finegan, T. A., & Siegfried, J.J. (2000). Are student ratings of teaching effectiveness influenced by instructors' English language proficiency? *The American Economist, 44*: 17-29.

Fisher, A. T., Alder, J. G., and Avasalu, M. W. (1998). Lecturing performance appraisal criteria: staff and student differences. *Australian Journal of Education ,42* (2): 153-168.

Ford, J.K., MacCallum, R.C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39* (2): 291-314.

Frey, P.W. (1973). Student ratings of teaching: Validity of several rating factors. *Science, 182*: 83-85.

Fuhrmann, B.S., and Grasha, A. F. (1983). *A Practical Handbook for College Teachers*. Boston: Little, Brown.

Fullan, (2001). *The new meaning of educational change* (3rd ed.). London: Routledge Falmer.

Gill, S., & Rebrova, A. (2001). Native and non-native: together we're worth more. *ELT Newsletter*, Article 52. Retrieved 3 April 2008 from http://www.eltnewsletter.com/back/March2001/art522001.htm

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, *15*: 1-13.

Glaser, B.G. (1978). *Theoretical sensitivity.* Mill Valley, California: The Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Chicago: Aldine.

Goodwin, L. D. & Stevens, E. A. (1993). The influences of gender on university faculty members' perceptions of 'good' teaching. *Journal of Higher Education*, *64*(2): 166-186.

Gorard, S., & Taylor, C. (2004). *Combining methods in educational research*. London: Open University.

Greene, J. C., Caracelli, V.J., & Graham, W. F. (1989). Toward a conceptual framework for mixed mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*(3): 255-274.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychology*, *52*(11): 1182-1186.

Greenwald, A.G., & Gillmore, G.M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*: 1209-1217.

Gregson, J. A. (1998). Editorial: reflecting on qualitative research and vocational education. *Journal of Vocational Education Research, 23* (4): 265-270.

Hammersley, M. (1996). Relationship between qualitative and quantitative research: Paradigm loyalty versus methodological eclecticism. In J.T.E. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences*. Leicester: BPS Books.

Harvey, L. (2003). Editorial: Student feedback. *Quality in Higher Education, 9*(1): 3-20.

Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: galloping polls in the 21st century. *Education policy Analysis Archives*, *5*(6): 1-32.

Hativa, N. (2000). *Teaching for effective learning in higher education*. Dordrecht: Kluwer Academic Publishers.

Hayton, G.E. (1983). An investigation of the applicability in technical and further education of a student evaluation of teaching instrument. Unpublished M.Ed. thesis, University of Sydney, Australia.

Hogan, T. P. (1973). Similarity of student ratings across instructors, courses, and time. *Research in Higher Education*, *1*: 149-154.

Holliday, A. (2002). *Doing and writing qualitative research*. London: Sage Publications.

Hong, L. K. (1984). List processing free responses: Analysis of open-ended questions with word processor. *Qualitative Sociology, 7* (1): 98-109.

Honig, H. (1997). Positions, power, and practice: Functionalist approaches and translation quality assessment. *Current Issues in Language and Sociology, 4* (1): 6-34

Howard, G. S., & Maxwell, S. E. (1980). Correlation Between Student Satisfaction and Grades: A Case of Mistaken Causation? *Journal of Educational Psychology*, *72*(6): 810-820.

Hutcheson, G. & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage.

Issan, S. A., & Gomaa, N.M. (2010). Post basic education reforms in Oman: A case study. *Literacy Information and Computer Education Journal (LICEJ), 1* (1). Retrieved 15 August 2010, from http://infonomics-society.org

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come [Electronic Version]. *Educational Researcher, 33* (7): 14-26. Retrieved 20th April, 2007 from http://www.jstor.org/stable/3700093

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, *39*: 31-36.

Kamindo, C. (2008). I*nstructional supervision in an era of change: Policy and practice in primary education in Keny*a. Unpublished doctoral dissertation, University of Durham, UK.

Kane, R., Sandretto, S., & Heath, C. (2004). An investigation into excellent tertiary teaching: Emphasising reflective practice. *Higher Education*, *47*: 283-310.

Kember, D. & Gow, L. (1994). Orientations to teaching and their effect on the quality of student learning. *Journal of Higher Education*, *65*: 58-74.

Kember, D., & Leung, D. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment and Evaluation in Higher Education, 33*(4): 341-353.

Kent, R., & Lee, M. (1999). Using the Internet for market research: A study of private trading on the Internet. *Journal of the Market Research Society*, *41* (4): 377–381.

Kierstead, D., D'Agostin, P., & Dill, W. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, *80*(3): 342-344.

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50*(3): 402-413.

Kishor, N. (1995). The effect of implicit theories on raters' inference in performance judgment: consequences for the validity of student ratings of instruction. *Research in Higher Education, 36* (2): 177-195.

Kittleson, M. J. (1995). An assessment of the response rate via the postal service and email. *Health Values, 18*(2): 27–29.

Koh, H. C., & Tan, T. M. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management*, *11*(4): 170-178.

Kolitch, E., & Dean, A.V. (1999). Student ratings of instruction in the USA: Hidden assumptions and missing conceptions about 'good' teaching. *Studies in Higher education*, *24*(1): 27-42. Retrieved 31 August, 2007 from http://dx.doi.org/10.1080/03075079912331380128

Kulik, J.A., & McKeachie, W.J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, *3*: 210-240.

Kwan, K. P. (2000). *How University Students Rate their Teachers: A Study of the rating attitudes and behaviours of university students in teaching evaluations*. Unpublished doctoral dissertation, University of Durham, UK.

L'Hommedieu, R.L., Menges, R. J., & Brinko, K. T.(1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*(2): 232-241.

Langbein, L.I. (1994). The validity of student evaluations of teaching. *Political Science and Politics*, September: 545-553.

Leckey, J., & Neville, N. (2001). Quantifying quality: the importance of student feedback. *Quality in Higher Education, 7*(1): 19-32.

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. *The International Journal of Educational Management*, *17* (1): 37-43.

Lin, W., Watkins, D., & Meng, Q. (1994). A cross-cultural investigation into students' evaluation of university teaching. *The Chinese University of Hong Kong (CUHK) Education Journal*, *22* (2): 291-304.

Lin, W., Watkins, D., & Meng, Q. (1995). Students' evaluations of university teaching: A Chinese perspective. *Higher Education Research and Development*, *14*(1): 61-74.

Loder, C. (1990). The introduction of staff appraisal in universities as a method of quality assurance. In C. loder (Ed.), *Quality assurance and accountability in higher education*. London: Kogan Page.

Lowman, J. (1995). *Mastering the techniques of teaching*. San Francisco: Jossey-Bass.

Ludwig, J.M., & Meacham, J.A. (1997). Teaching controversial courses: Student evaluations of instructor and content. *Educational Research Quarterly, 21*(1): 27-38.

Maamouri, M. (1998). *Arabic diglossia and its impact on the quality of education in the Gulf region.* A paper presented at Language Education and Human Development: Moving Forward Workshop, Mediterranean Development Forum, Sep. 3-6, Marrakech, Morocco.

MacCallum, R.C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1): 84-99.

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin & Associates, *Changing Practices in Evaluating Teaching*. Bolton, MA: Anker Publishing Company.

Marques, T. E., Lane, D. M., and Dorfman, P. W. (1979). Toward the development of a system for instructional evaluation: Is there consensus regarding what constitutes effective teaching? *Journal of Educational Psychology*, *71* (6): 840-849.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in the evaluation of teaching. *American Educational Research Journal*, *17*: 219-237.

Marsh, H.W. (1981). Students' evaluations of tertiary instruction: testing the applicability of American surveys in an Australian setting. *Australian Journal of Education*, *25*: 177-192.

Marsh, H.W. (1982a). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*: 77-95.

Marsh, H. W. (1982b). The use of path analysis to estimate teacher and course effects in student rating's of instructional effectiveness. *Applied Psychological Measurement*, *6*: 47-59.

Marsh, H.W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*: 150-166.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*(5): 707-754.

Marsh, H.W. (1986). Applicability paradigm: students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology*, *78* (6): 465-473.

Marsh, H.W. (1987). Student evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*: 253-388.

Marsh, H.W. (2001). Distinguishing between good (useful) and bad workload on students' evaluations of teaching. *American Educational Research Journal, 38*(1): 183-212.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R.P. Perry and J.C. Smart (Eds.), *The Scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). The Netherlands: Springer.

Marsh, H.W., & Cheng, J. (2008). *The national student survey: A multilevel analysis of discipline effects*. A paper presented at the The NSS Conference, 8th May 2008, Nottingham, UK.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluation of university teaching: A multi-dimensional perspective. In: J.C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (pp. 143-232). New York: Agathon Press.

Marsh, H. W., & Dunkin, M.J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry and J.C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241-320). New York: Agathon Press.

Marsh, H.W., Hau, K., Chung, C., & Siu, T. (1998). Confirmatory factor analysis of Chinese students' evaluations of university teaching. *Structural equation Modeling*, 5(2): 143-164.

Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of students' evaluations of college training. *American Educational Research Journal*, *21*: 341-366.

Marsh, H. W., & Hocevar, D. (1991a). Student evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education*, *7*(4): 303-314.

Marsh, H. W., & Hocevar, D. (1991b). The multidimensionality of students' evaluations of teaching effectives: The generality of factor structure across academic discipline, instructor level and course level. *Teaching and Teacher Education*, *7* (1): 9-18.

Marsh, H.W., and Overall, J.U. (1979). Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 10*: 139-147.

Marsh, H.W. & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, *30* (1): 217-251.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52* (11): 1187-1197.

Marsh, H. W., and Roche, L. A. (April, 1998). Effects of grading leniency and low workloads on students' evaluations of teaching. Paper presented at the 79th Annual Meeting of the American Educational Research Association, San Diego, California.

Marsh, H.W., & Roche, L.A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? J*ournal of Educational Psychology, 92*: 202-228.

Marsh, H. W., Touron, J., & Wheeler, B. (1985). Students' evaluations of university instructors: The applicability of American instruments in a Spanish setting. *Teaching & Teacher Education*, *1* (2): 123-138.

Marsh, H.W., and Ware, J.E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox Effect. *Journal of Educational Psychology, 74*: 126-134.

Martin, J. R. (1998). Evaluating faculty based on student opinions: problems, implications, and recommendations from Deming's theory of management perspective. *Issues in Accounting Education*, *13*(4): 1079-1094.

Matsuda, A. & Matsuda, P. K. (2001). Autonomy and collaboration in teacher education: Journal sharing among native and non-native English-speaking teachers. *The CATESOL Journal, 13* (1): 109-121.

Maum, R. (2002). Non-native English speaking teachers in the English teaching profession. *Eric Digest.* Retrieved 11 June 2009 from http://www.cal.org/ericcll/digest/0209maum.html.

McCarger, D. F. (1993). Teacher and student role expectations: Cross-cultural differences and implications. *The Modern Language Journal, 77*: 192-207.

McKeachie, W. J. (1979). Student Ratings of Faculty: A Reprise. *Academe, 65*: 384-397.

McKeachie, W.J.(1994). *Teaching tips: Strategies, research, and theory for college and university teachers* (9th ed.). Lexington: D.C. Heath & Company.

McKeachie, W. J. (1997a). Good teaching makes a difference- and we know what it is. In: R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 396-408). New York: Agathon Press.

McKeachie, W. J. (1997b). Student ratings: The validity of use. *American Psychologist*, *52*(11): 1218-1225.

McMahon, S.R., Iwamoto, M., Massoudi, M.S., Yusuf, H.R., Stevenson, J.M., David, F., Chu, S.Y., and Pickering, L.K. (2003). Comparison of e-mail, fax, and postal surveys of pediatricians. *Pediatrics, 111*(4): 299-303.

Medgyes, P. (1992). Native or non-native: Who's worth more? *ELT Journal, 46* (4): 340-349.

Mehta, R., & Sivadas, E. (1995). Comparing the response rates and content in mail versus electronic mail surveys. *Journal of the Market Research Society, 37*(4): 429-439.

Meirovich, G., & Romar, E. J. (2006). The difficulty in implementing TQM in higher education instruction: The duality of instructor/student roles. *Quality Assurance in Education*, *14*(4): 324-337.

Meleis, A. I. (1982). Arab students in western universities: Social properties and dilemmas. *The Journal of Higher Education, 53*(4): 439-447.

Mercer, J. E. (2004). *The impact of faculty appraisal at tertiary level: Two exploratory case studies.* Unpublished EDD dissertation, Open University, UK.

Miller, M.B. (1995). Coefficient alpha: a basic introduction from the perspective of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*(3): 255-273.

Mitchell, D. E., & Kerchner, C. T. (1983). Collective bargaining and teacher policy. In L. S. Shulman & G. Sykes (Eds.), *Handbook of teaching and policy*. New York: Longman.

Morgan, D. L. (1998). Practical strategies for combining qualitative and quantitative methods: Applications to health research. *Qualitative Health Research,* 8(3): 362-376.

Morgan, J. & Davies, T. (2006). Analysis of bias in student evaluations of faculty at an all female Arab university in the Middle East. *Learning and Teaching in Higher Education: Gulf Perspectives, 3*(2). Retrieved 14 April 2009 from http//www.zu.ac.ae/lthe/lthe03_02_02morgan.htm

Moritsch, B.G., & Suter, W.N. (1988). Correlates of halo error in teacher evaluation. *Educational Research Quarterly, 12*(3): 29-34.

Morreale, J.C. (1999). Post tenure review: Evaluating teaching. In P. Seldin & Associates (Eds.), *Changing Practices in Evaluating Teaching*. Bolton, MA: Anker Publishing Company.

Moussu, L. (2000). Native versus non-native speakers of English: Students' reactions. Retrieved 16 August 2008 from http://www.moussu.net/courses/portfolio/540.pdf.

Murray, H. G. (1983). Low inference classroom teaching behaviours and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, *71*: 856-865.

Murray, H. G. (1987). *Impact of student instructional ratings on quality of teaching in higher education*. Paper presented at the 1987 annual meeting of the American Educational Research Association, Washington, D. C.

Murray, H. G. (1991). Effective teaching behaviors in the college classroom. In: J.C. Smart (Ed.), *Higher Education: Handbook of Theory and Research*. New York: Agathon Press.

Murray, H.G., Rushton, P. J., & Paunonen, S. V. (1990) Teacher personality traits and student instructional ratings in six types of university courses. *Journal of educational psychology*, *82*: 250-261.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor X Lecture: A Paradigm of Educational Seduction. *Journal of Medical Education*, *48*: 630-635.

Nayar, P. B. (1994). Whose English is it?. *TESL-EJ, 1*, 1. Retrieved 23 September 2009 from http://www.kyoto-su.ac.jp/information/tesl-ej/ej01/f.1.html.

North, J. D. (1999). Administrative courage to evaluate the complexities of teaching. In P. Seldin & Associates (Eds.), *Changing Practices in Evaluating Teaching*. Bolton, MA: Anker Publishing Company.

Nunnally, J.O. (1978). *Psychometric theory*. New York: McGraw-Hill.

Ogier, J. (2003). *Evaluating student rating of teaching form*. A paper presented at the Evaluation and Assessment: A Commitment to Quality Conference, Adelaide, New Zealand (November).

Okpala, C. O., & Ellis, R. (2005). The perceptions of college students on teacher quality : A focus on teacher qualifications. *Education*, *126*: 374-378.

Oman Accreditation Council (2005). *Executive Summary: Requirements for Oman's System of Quality Assurance in Higher Education*. Muscat, Oman.

Oman Accreditation Council (2006). *Plan for an Omani higher education quality management system ("The Quality Plan").* A joint initiative of the Ministry of Higher Education and the Oman Accreditation Council, Muscat, Oman.

Oman Accreditation Council (2007). *Oman academic standards for General Foundation Programs.* Muscat, Oman.

Oman Quality Network (2008). *Proceedings of the Oman national quality conference: Quality Management and Enhancement in Higher Education*, 29-29th October 2008. Muscat, Oman.

Onwuegbuzie, A. J., & Daniel, L. G. (2002). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development*, *35*: 89-103.

Onwuegbuzie, A. J., & Daniel, L. G. (2004). Reliability generalization : The importance of considering sample specificity, confidence intervals, and subgroup differences. *Research in the Schools*, *11*(1): 61-72.

Onwuegbuzie, A.J., Witcher, A.E., Collins, K.M., Filer, J.D., Wiedmaier, C.D., Moore, C.W. (2007). Students' perceptions of characteristics of effective college

teachers: A Validity study of a teaching evaluation form using a mixed-method analysis. *American Educational Research Journal*, *44*(1): 113-160.

Oppermann, M. (1995). E-mail surveys—Potentials and pitfalls. *Marketing Research, 2*(3): 28–33.

Ory, J.C. & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research, 109*: 27-44.

Overall, J. U., & Marsh, H. W. (1979). Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology, 71*: 856-865.

Overall, J. U., & March, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, *72*: 321-325.

Overton, J., & van Dierman, P. (2003). Using quantitative techniques. In R. Scheyvens & D. Storey(Eds.), *Development fieldwork: A practical guide* (pp. 37–56). Thousand Oaks, CA: Sage

Pallant, J. (2007). *SPSS survival manual.* Berkshire: Open University Press.

Parker, L. (1992, July). Collecting data the e-mail way. *Training & Development*, *46*(7): 52-54.

Parker, O. D. (1986). Cultural clues to the Middle Eastern student. In Valdes, J. M. (Ed.), *Culture Bound: Bridging the cultural gap in language teaching* (pp. 94-101). New York: Cambridge University Press.

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment & Evaluation in Higher Education*, *23*(2): 165-178.

Pennington, M. C., & Young, A. L. (1989). Approaches to faculty evaluation for ESL. *TESOL Quarterly*, *23*(4): 619-646.

Penny, A. R. (2003). Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education*, *8*(3): 399-411.

Penny, A. R. (2004). Effects of student ratings feedback and a group intervention on the quality of university teaching: A randomised controlled trial. Unpublished doctoral dissertation, University of Durham, UK.

Peterson, K. D. (1995). *Teacher Evaluation: A comprehensive guide to new directions and practices*. California: Corwin Press.

Petridou, E., & Sarri, K. (2004). Evaluation research in business schools: Students' rating myth. *The International Journal of Educational Management*, *18*(3): 152-159.

Phillipson, R. (1992). *Linguistic Imperialism.* Oxford: Oxford University Press.

Phillipson, R. (1996). ELT: The native speaker's burden. In T. Hedge & N. Whitney (Eds.), *Power, pedagogy & practice* (pp.23-30). Oxford: Oxford University Press.

Pratt D. D.; Kelly M.; Wong W. S. S. (1999). Chinese conceptions of 'effective teaching' in Hong Kong: towards culturally sensitive evaluation of teaching. *International Journal of Lifelong Education, 18* (4): 241-258.

Pring, R. (2001). The virtues and vices of an educational researcher [Electronic Version]. *Journal of Philosophy of Education, 35*(3): 407-421. Retrieved 15th February, 2008 from http://www.blackwell-synergy.com/toc/jope/35/3

Radford, A. E. (1980). Outstanding teacher characteristics as perceived by Saudi Arabian ESL students and American college students. Unpublished MA thesis, University of California, USA.

Rampton, M.B.H. (1996). Displacing the native speaker: Expertise, affiliation, and inheritance. In T. Hedge & N. Whitney (Eds.), *Power, pedagogy & practice* (pp.9-22). Oxford: Oxford University Press.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.

Ramsden, P. (2003). *Learning to teach in higher education* (2nd ed.). London: Routledge.

Ranchhod, A., & Zhou, F. (2001) Comparing respondents of e-mail and mail surveys: understanding the implications of technology. *Marketing Intelligence & Planning, 19* (4): 254 – 262.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2010). *MLwiN Version 2.17*. Centre for Multilevel Modelling, University of Bristol, UK.

Rashdall, H. (1936) *The Universities of Europe in the Middle Ages*. London: Oxford University Press.

Raykov, T, & Shrout, P.E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9*: 195-212.

Raymond, S. M. (2001). Excellent teaching: perceptions of Arab, Chinese, and Canadian students. In *Voices of Arabia* (pp.17-30), Sheffield: University of Sheffield.

Raymond, S. M. (2008). Effective and ineffective university teaching from the students' and faculty's perspectives: Matched or mismatched expectations? Unpublished doctoral dissertation, University of Exeter, UK.

Reid, L.D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*(3): 137–152.

Richards, J.C., & Rodgers, T.S. (2001). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Richardson, J.T. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment and Evaluation in Higher Education, 30*(4): 387-415.

Richardson, K. E. (1998). Quantifiable feedback: can it really measure quality? *Quality Assurance in Education*, *6*(4): 212-219.

Roche, L.A. & Marsh, H.W. (2002). Teaching self-concept in higher education: reflecting on multiple dimensions of teaching effectiveness. In N. Hativa & P. Goodyear (Eds.), *Teacher thinking, beliefs, and knowledge in higher education.* Dordrecht: Kluwer.

Rowley, J. (2003). Designing student feedback questionnaires. *Quality Assurance in Education*, *11*(3): 142-149.

Russell, A. (2004). Zayed University students' teaching and learning beliefs and preferences: An analysis based on the surface versus deep learning approach. *Learning and Teaching in Higher Education: Gulf Perspectives*, *1*. Retrieved 12 September, 2008 from http://www.zu.ac.ae

Ryan, J. J., Anderson, J. A., & Birchler, A. B. (1980). Student evaluations: The faculty responds. *Research in Higher Education*, *12*: 317-333.

Saafin, S.M. (2005). *An investigation into Arab students' perceptions of effective EFL teachers at university level.* Unpublished doctoral dissertation, University of Exeter, UK.

Saafin, S.M. (2008). Arab tertiary students' perceptions of effective teachers. *Learning and Teaching in Higher Education: Gulf Perspectives*, *5*(2). Retrieved 12 September, 2008 from http://www.zu.ac.ae

Sacks, P. (1996). *Generation X Goes to College*. Chicago: Open Court.

Salomon, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher*, August-September: 10-18.

Schaefer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public Opinion Quarterly, 62*(3): 378-397.

Schaeffer, G., Epting, K., Zinn, T., & Buskit, W. (2003). Student and faculty perceptions of effective teaching: A successful replication. *Teaching of Psychology, 30*: 133-136.

Schuldt, B. A., & Totten, J. W. (1994). Electronic mail vs. mail survey response rates. *Marketing Research, 6*(1): 36-39.

Scriven, M. (1981). Summative teacher evaluation. In J. Millan (Ed.) *Handbook of teacher evaluation* (pp.244-271). Beverly Hills: Sage Publication.

Scriven, M. (1987). Validity in Personnel Evaluation. *Journal of Personnel Evaluation in Education*, *1*: 9-23.

Scriven, M. (1988). The validity of student ratings. *Instructional Evaluation*, *9*(2): 5-18.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment, Research & Evaluation*, *4*(7). Retrieved 30 August, 2007 from http://PAREonline.net/getvn.asp?v=4&n=7

Seldin, P. (1993a). *Improving and evaluating teaching.* Paper presented at the American Council on Education Department Chairs Seminar, Washington, DC.

Seldin, P. (1993b). The Use and Abuse of Student Ratings of Instruction. *The Chronicle of Higher Education*, *39*, 1-40.

Seldin, P. (1998). *The teaching portfolio*. Paper presented for the American Council on Education, Department Chairs Seminar, San Diego, CA.

Seldin, P. (1999). Current practices –good and bad- nationally. In P. Seldin & Associates (Eds.), *Changing Practices in Evaluating Teaching*. Bolton, MA: Anker Publishing Company.

Shermis, M. D., & Lombard, D. (1999). A comparison of survey data collected by regular mail and electronic mail questionnaires. *Journal of Business and Psychology, 14*(2): 341-354.

Shih, T., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20* (3): 249-271.

Sidanius, J., & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology, 19*: 174–197. Retrieved 8 July 2010 from http://www.jstor.org/stable/4132809.

Sixbury, G.R. & Cashin, W.E. (1995) IDEA technical report No. 9: *Descriptions of database for the IDEA diagnostic form*. Publication of the Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University.

Small, R. (2001). Codes are not enough: What philosophy can contribute to the ethics of educational research [Electronic Version]. *Journal of Philosophy of Education*, *35*(3): 387-405. Retrieved 15th February, 2008 from http://www.blackwell-synergy.com/toc/jope/35/3

Smith, L. (2006). Teachers' conceptions of teaching at a Gulf university: A starting point for revising a teacher development program. *Learning and Teaching in Higher Education: Gulf Perspectives*, *3* (1). Retrieved 12 September, 2008 from http://www.zu.ac.ae

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. R. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching, 50* (2): 44-49.

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives*, *8*(50). Retrieved 21 September, 2007 from http://epaa.asu.edu/epaa/v8n50.html.

Sproull, L. S. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal, 29*(1): 159-169.

Stevens, G. E. (1978). Teaching by whose objectives? The view of students and teachers. Unpublished manuscript.

Stevens, G. E., and Marquette, R. P. (1979). Differing student and faculty perceptions of teaching effectiveness and the value of student evaluations. *POD Quarterly*, *1* (4): 207-219.

Stratham, A., Richardson, L., & Cook, J.(1991). *Gender and university teaching: A negotiated difference.* Albany: Suny Press.

Tabachnick, B. G. & Fidell, L.S. (2001). *Using multivariate statistics* (4th ed.). Boston: Pearson Education.

Tabachnick, B. G. & Fidell, L.S. (2007). *Using multivariate statistic*s (5th ed.). Boston: Pearson Education.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oakes, CA: Sage Publications.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. London: Sage Publications.

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research & Development in Education, 28*: 169–173.

Teddlie, C., Stringfield, S., & Burdett, J. (2003). International Comparisons of the Relationships among Educational Effectiveness, Evaluation and Improvement Variables: An Overview. *Journal of Personnel Evaluation in Education*, *17*(1): 5-20.

Thach, E. (1995). Using electronic mail to conduct survey research. *Educational Technology*, March-April: 27-31.

Thakkar, J., Deshmukh, S. G., & Shastree, A. (2006). Total quality management (TQM) in self-financed technical institutions: A quality function deployment (QFD) and force field analysis approach. *Quality Assurance in Education*, *14*(1): 54-74.

Theall, M., Franklin, J., (2001). Looking for bias in all the wrong places: a search for truth or a witch hunt in student ratings of instruction. In: Theall, M., Abrami, P., Mets, L. (Eds.), *The student ratings debate: Are they valid? How can we best use them?* (pp. 45–56). San Francisco: Jossey-Bass.

Trout, P. A. (1997). What the numbers means: Providing a context for numerical student evaluations of courses. *Change*, September/ October: 24-30.

Tse, A. C. B. (1998). Comparing the response rate, response speed and response quality of two methods of sending questionnaires: e-mail vs. mail, *Journal of the Market Research Society, 40*(4), 353-361.

Tse, A. C. B., Tse, K. C., Yin, C. H., Ting, C. B., Yi, K. W., Yee, K. P., & Hong, W. C. (1995). Comparing two methods of sending out questionnaires: E-mail versus mail. *Journal of the Market Research Society, 37*(4): 441-446.

Ustunluoglu, E. (2007). University students' perceptions of native and non-native teachers. *Teachers and Teaching: Theory and Practice*, *13* (1): 63-79.

Valdes, J. M. (Ed.) (1986). *Culture Bound: Bridging the cultural gap in language teaching*. New York: Cambridge University Press.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher education*, *23* (2): 191-211.

Wajnryb, R. (1990). *Grammar dictation*. Oxford: Oxford University Press.

Walumbwa, F. O., & Ojode, L. A. (2000). *Gender stereotype and instructors' leadership behavior: Transformational and transactional leadership*. Midwest Academy of Management Annual Conference at Chicago, March 30th-April 1st.

Warwick, D. P. and Osherson, S. (1973). Comparative analysis in the social sciences. In D. P. Warwick and S. Osherson (Eds.), *Comparative Research Methods: An Overview*. Englewood Cliffs, NJ: Prentice-Hall.

Watkins, D. (1992). Evaluating the effectiveness of tertiary teaching: A Hong Kong perspective. *Education Research Journal*, *7*: 60-67.

Watkins, D. (1994). Student evaluations of university teaching: A cross-cultural perspective. *Research in Higher Education*, *35* (2): 251-266.

Watkins, D. & Akande, A. (1992). Student evaluations of teaching effectiveness: a Nigerian investigation. *Higher Education*, *24* (4): 453-463.

Watkins, D. & Gerong, A. (1992). Evaluating tertiary teaching: a Filipino investigation. *Educational and Psychological Measurement*, *52*: 727-734.

Watkins, D. & Regmi, M. (1992). Student evaluations of tertiary teaching: a Nepalese investigation. *Educational Psychology*, *12* (2): 131-142.

Watkins, D., & Thomas, B. (1991). Assessing teaching effectiveness: An Indian perspective. *Assessment and Evaluation in Higher Education*, *16* (3): 185-198.

Watkins, D., Marsh, H.W., & Young, D. (1987). Evaluating tertiary teaching: A New Zealand perspective. *Teaching & Teacher Education*, *3* (1): 41-53.

Weible, R., & Wallace, J. (1998). Cyber research: The impact of the Internet on data collection. *Marketing Research*, *10*: 19-25.

Wiles, R., Charles, V., Crow, G. & Heath, S. (2006). Research ethics and data quality: The implications of informed consent. *International Journal of Social Research Methodology, 19*(2): 83-95.

Williams. W. M., & Ceci, S. J. (1997). How'm I Doing? Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning*, *29* (Sept./ Oct.): 12-23.

Witcher. A. E., Onwuegbuzie, A.J., & Minor, L. C. (2001). Characteristics of effective teachers: Perceptions of pre-service teachers. *Research in the Schools, 8*(2): 45-57.

Wotruba, T. R., and Wright, P. L. (1975). How to develop a teacher-rating instrument: a research approach. *Journal of Higher Education*, *46* (6): 653-663.

Young, P., Delli, D. A., & Johnson, L. (1999). Student evaluation of faculty: Effects of purpose on pattern. *Journal of Personnel Evaluation*, *13* (2): 179-190.