An investigation into the oral English language proficiency gain of pupils taught by native English-speaking teachers in Hong Kong secondary schools

Gray, Jeremy

AN INVESTIGATION INTO THE ORAL ENGLISH LANGUAGE PROFICIENCY GAIN OF PUPILS TAUGHT BY NATIVE ENGLISH-SPEAKING TEACHERS IN HONG KONG SECONDARY SCHOOLS

Ed. D.

JEREMY GRAY

UNIVERSITY OF DURHAM

2002

An Investigation into the Oral English Language Proficiency Gain of Pupils Taught by Native English-speaking Teachers in Hong Kong Secondary Schools.

A Thesis by

Jeremy Gray

Supervisor: Professor Peter Tymms

Submitted for the Degree of

Doctor of Education

in the

School of Education of the University of Durham

December 2002

Abstract of thesis entitled "An Investigation into the Oral English Language Proficiency Gain of Pupils Taught by Native English-speaking Teachers in Hong Kong Secondary Schools"

Submitted by Jeremy Gray

for the degree of Doctor of Education

At the University of Durham in December, 2002

ABSTRACT

Purpose

This study examines the extent to which Native English-speaking Teachers (NETs) have an impact on the oral English language proficiency gain of pupils taught in secondary schools in Hong Kong i.e. the so-called 'NET effect'. The principal aim was to determine whether the oral proficiency gain of subjects involved in this study was greater in students taught by NET teachers than it was in students taught by local teachers. Through the examination of Time one and Time two oral assessment data, this study also sets out to investigate the nature and strength of other predictor variables for the outcome variable 'Time two oral assessment'. Through a number of different statistical modelling techniques this study also sought to establish the model that would account for or 'explain' as much variance as possible between the Time one and Time two assessment scores.

Procedures

A randomised, stratified sample of secondary schools that was representative of the whole population secondary students in Hong Kong who are studying English was generated. From this sample, one thousand four hundred and twenty four students from forms one, three and four were selected and an English language oral proficiency test, specifically developed for this study was administered as a pre and post test. The period of data collection was a two year period, from the beginning of the 1998-1999 academic year to the end of the 1999-2000 academic year.

The oral proficiency assessment instrument was designed, and piloted by a small team of trained assessors, and a standardised procedure was established for conducting the assessments. Hong Kong NET and local teachers were trained in the procedures and use of materials and techniques required to administer the assessments in specifically dedicated language assessment workshops. The assessments were then administered by the trained group of teachers who taped all of the interviews to allow monitoring to take place and to provide a data source for a second stage interview analysis (not covered in this thesis).

The resulting pre-test and post-test data was then analysed through the use of a number of statistical techniques. In the first instance, a descriptive analysis was conducted in order to satisfy the assumptions on which traditional statistical analysis is based. The data analysis then proceeded with a number of scaling processes and was finally analysed to determine whether or not any significant 'NET effect' had been

detected. In addition, the analysis also considered whether any of the other variables could be considered good predictors of the final post-test score.

Major Findings

Analysis of data produced from the Time one and Time two oral assessments revealed a number of important findings. Students did make significant oral English language proficiency gain as measured by the specially developed instrument. This gain was significant regardless of whether the students' results were analysed by whole sample or by separate year/age group. An analysis of means revealed that on average, the mean scores of students attending EMI schools were significantly higher than those attending CMI schools indicating that the medium of instruction is potentially a strong predictor of the Time two assessment score. In the post test analysis of means, students taught by NETs performed better than those taught by 'both' [NETs *and* local teachers] and in general, students taught by local teachers also performed better than those taught by 'both'. There was little difference between the scores of students taught by NETs and local teachers.

The banding of the schools was also found to be an important predictor variable, with the average scores of students in high band schools significantly higher than their peers in medium and low band schools.

Multiple regression analysis also revealed some important findings. When the modelling was conducted on the whole sample, the medium of instruction, the school level and NET teacher were all found to be significant predictor variables although in the case of the latter, the effect was small. When modelled by separate form/age group,

similar results were found with Form one and with Form three students and again the medium of instruction, the school level and NET teacher were significant predictor variables although in the case of NET teacher, the effect was again small.

The findings of this thesis suggest that in terms of measuring value-added between Time one and Time two, there are indeed strong predictor variables such as medium of instruction, school level and student level. However, in trying to evaluate the contribution of NETs to students' oral English language proficiency gain over a two-year period, there is some evidence of a so-called 'NET effect' although this is rather weak, suggesting that more research is required to investigate this question more thoroughly.

ACKNOWLEDGEMENTS

The author wishes to acknowledge and thank the many people who have made a contribution towards the background research, preparation of assessment instruments, data collation and preparation of this thesis.

Thanks go firstly to my supervisor Professor Peter Tymms, for his encouragement, insights and careful personal, professional and technical guidance. His comments and suggestions were much appreciated and have made a positive contribution towards this thesis.

The author also wishes to thank all the members of the MENETS project whose hard work and continuous encouragement have been very much appreciated. Particular thanks to Dr Peter Storey, the MENETS project director and to the key members of the project team Ms Elena Wang, Ms Jasmine Luk, Dr Rita Berry and Dr Angel Lin. Also, sincere thanks go to the research assistants who worked on the MENETS project at different stages, especially Anna, Ted, Cynthia, Iddy and Stephen.

The author would also like to thank all those friends and colleagues within the Hong Kong Institute of Education for their continuous support, comments and constructive advice, especially Dr Vernon Crew and other members of the Department of English.

Many NET and local teachers throughout Hong Kong were at different stages involved in the MENETS project and have given considerable time and effort to help make the project a success. Thank you all.

Finally, love and deepest thanks go to my long-suffering family, Chris, Lara and Naomi for their love, support and understanding.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 7

Chapter 9

LIST OF FIGURES

Chapter 7

LIST OF ABBREVIATIONS

CMI	Chinese as a Medium of Instruction
ED	Education Department
ELT	English Language Teaching
EELTS	Expatriate English Language Teacher Scheme
EELTMS	Expatriate English Language Teacher's Modified Scheme
ECR(1)	(First) Education Commission Report
EFL	English as a Foreign Language
EMI	English as a Medium of Instruction
ESL	English as a Second Language
HKEA	Hong Kong Examinations Authority
HKIEd	Hong Kong Institute of Education
ILE	Institute for Language in Education
L2	Second language
LAD	Language Acquisition Device (Chomsky)
MENETS	Monitoring and Evaluation of the Native-speaking English Teacher Scheme
MOI	Medium of Instruction
NET	Native-speaking English Teacher
SCOLAR	Hong Kong Standing Committee on Language Education and Research
SLA	Second Language Acquisition
TOC	Target Oriented Curriculum

CHAPTER ONE

INTRODUCTION

This chapter presents the background and context in which the present study is set by firstly presenting the historical and socio-political framework within which the NET scheme was established and still exists today in Hong Kong. Following this, the chapter then presents the rationale and the need underlying studies such as this thesis which investigate different aspects related to the effectiveness of Native English-speaking Teachers (NETs) in Hong Kong.

I. BACKGROUND TO THE STUDY

1. Origins of the Current NET Scheme in Hong Kong

> To make an immediate impact on improving the English language standard of our students, we will implement a new Native-speaking English Teachers Scheme, providing more than 700 additional native-speaking English teachers for secondary schools from next year (Tung Chee Hua, 1997).

Thus from Mr. Tung Chee Hua's (the Chief Executive of Hong Kong SAR) first policy address in October 1997, the current NET scheme was born, and was first implemented in September 1998. In the 1998-99 school year, 338 NET teachers were recruited with the number increasing to 440 by the 1999-2000 school year. The number currently stands at 441. In addition to these measures, two separate school organisations

received funding from two bodies established to improve educational standards, namely The Hong Kong Language Fund and The Quality Education Fund. The funding was aimed at introducing expatriate teachers into primary schools in 1998-2000 and under these two schemes, 16 teachers were recruited. Four years later, in his 2001 policy address, Mr. Tung Chee Hua announced the expansion of the NET scheme to potentially include all primary schools in Hong Kong as well as continue with the scheme, as it exists in secondary schools (i.e. nearly all):

> Since the 1998-1999 school year, the Native-speaking English Teacher Scheme has been operating in secondary schools. Although the scheme encountered some teething problems, it has gradually brought about a new culture of English language teaching in our schools that is widely supported. For effective language learning, it should start as early as possible. From the start of the next school year, we will strengthen English language teaching in primary schools with various initiatives. Our targets include providing native English-speaking teachers or teaching assistants in every primary school and the organisation of more extra-curricular activities using English. (Tung Chee Hua 2001)

There seemed to be a general acceptance by the public that the deployment of NETs would necessarily bring about an improvement of English language standards in the classroom. Indeed, there has been little if any public debate or criticism of the implicit linkage between the deployment of native speakers and improved language proficiency. Neither of the two speeches (above) contains any evidence, explanation or justification to substantiate this link. Furthermore, there was no consultation between the HK government and those working in the field of education to gauge opinion on this

issue. It would seem that there was an appeal to the 'common sense' of the layperson, based on the principle that of course English language levels can be increased by the deployment of native speakers. Since the English level of NET teachers is better than that of local teachers, all else naturally follows it might be argued. However, this assertion gives rise to a number of important questions. Firstly, what are the theoretical and methodological rationales underlying the NET scheme? Secondly, are such models effective, since if they are not then the validity of such a scheme is fundamentally put into question. We will examine these two key questions by looking at the related literature

2. English language teaching in Hong Kong: a historical perspective

> Hong Kong is an international business, financial and trading center. English therefore has an important place in the economic life of our community. In order to maintain Hong Kong's international position, we have to ensure that we produce sufficient well-educated people able to communicate in both English and Chinese (ECR4 p.101).

In his chronology of English language teaching and learning in Hong Kong, Sweeting (1990, 1992) documents the use of English as a Medium of Instruction (EMI) in Hong Kong schools, beginning in as early as 1843. Prior to this, Hong Kong was little more than a fishing village, with migration of people into Hong Kong from uncolonised China only really beginning in the early 1840's. The driving force behind the early history of native-speaking English teachers in Hong Kong had more to do with missionary zeal and supporting the political and economic needs of the British Empire

than with establishing a sound educational or linguistic base for the local inhabitants. There was also a feeling by the British colonists that the locals had to be 'civilised' in the sense of freeing them from the shortcomings of Chinese educational traditions and systems. An educational conference in 1845 recommended that "...the study of English should in this English colony be encouraged as much as possible" Sweeting (1990:147) and in 1853 English teaching was introduced into two of the government aided schools under 'pupil-teachers' from the St. Paul's school. A conference in 1878 acknowledged the increasing demand for English language interpreters and endorsed the use of EMI in order to meet this demand. In response to the increased demand for English language teachers, a 'Normal School' was established in Wan Chai for which 10 students were selected to be trained as English language teachers. These early attempts at introducing English to the school curriculum seem to have had little effect however, since in the same year John Pope Hennessy[1], a former Governor of Hong Kong, observed that in Central School:

> "...during the whole year we have had 610 pupils attending this school ... and I asked how many of these were able to speak English and he said under 50 or 60, and this small number very imperfectly. We must not be satisfied with 60 out of 600 being able to speak English in our principal Government school." (Sweeting 1990:350).

Thus to the extent to which a 'language policy' existed, the emphasis was on the teaching of English first and foremost but this resulted in only a small minority of pupils acquiring English language skills to any usable level. As recently as 1917, another Governor, Sir Henry May, noted in one of his dispatches that:

> "the Government schools are almost exclusively for English and other 'foreign' children for Chinese and for Indians. They mainly give instruction in English, but there are Vernacular Grant Schools where Chinese is the sole language ... there is however, a strong demand among Chinese for more Vernacular education for the very poorest class of people ... I am of the opinion that such provisions cannot be conveniently made by the existing machinery. And new machinery must be sought." (Sweeting 1990:369).

The Burney Report of 1935 marked somewhat of a sea change in language education in Hong Kong. Not only did this report for the first time suggest that a Director of Education be appointed in order to provide some kind of effective educational leadership, but the emphasis moved for the first time away from English as being the main medium of instruction. Burney (op cit) notes that "Education policy in the colony should be gradually reoriented so as eventually to secure for the pupils, first a command of their own language sufficient for all needs of thought and expression, and secondly, a command of English limited to the satisfaction of vocational demands." (p.25). Burney's (1935) recommendations seem to have had some impact since Sweeting (1990) reports that by 1939 Cantonese was being used as the medium of instruction for subjects other than English in some of the lower classes of Anglo-Chinese schools. The Burney Report (1935) was thus the first time that the Hong Kong government acknowledged the importance of giving priority to the provision of vernacular primary education although little was done in terms of educational reform and the status quo remained until the next important landmark in policy reform in language in education in Hong Kong. It should be noted however that even this landmark report has been criticised as colonialist because in

its recommendations it stated that local (Hong Kong) students should learn only 'a command of English limited to the satisfaction of vocational demands' (Burney 1935:25). It was felt that this recommendation treated Hong Kong Chinese citizens as inferiors by depriving them of full rights to English culture (Sweeting (1992) citing Bentley (1988)).

The 1965 Cultural Revolution in China naturally had a considerable impact in Hong Kong. Sweeting (1992) notes that from about 1966 onwards, 'colonialism has been used as a pejorative slogan or ubiquitous scapegoat at a time when there has been little trace of colonization as a policy determinant' (p46). The riots of 1966 and 1967 were further manifestations of political, social and educational discontent and the establishment of the Students' Movement in 1971 and the Chinese Language Movement quickened the pace towards the vernacularisation of the curriculum.

A visiting group of educational experts led by Lord Llewellyn was invited to Hong Kong in 1982 to review Hong Kong's educational system and it was the findings of this Panel that really laid the foundations of Hong Kong's current 'Mother Tongue' policy. The importance of the English language to the future economic development of the territory was emphasised with the assertion that 'Hong Kong cannot afford to reduce the emphasis in its schools' (p28) but that 'the necessity for most students to learn two languages – English and Chinese – is an unusual privilege and burden' (p25). The Llewellyn Report (1982) came up with two key findings that should be highlighted, namely:

1. The medium of instruction in schools.

Llewellyn (1982) was hard-hitting and explicit in his views of "the present lamentable situation concerning the use of English as a medium of instruction" which he pointed out was likely to continue because of the basic issue "of whether it is possible to use a second language successfully as the vehicle for providing universal (compulsory) education in what is de facto, although not de jure, still a monolingual society as far as the vast majority of the population is concerned" (p26). The report further recognised the inevitability under the prevailing conditions of rote learning. In sum, there was a need to 'accept as a fact that the mother tongue is, all other things being equal, the best medium of teaching and learning' (Llewellyn 1982:28). The problem then, as it remains today, was the resistance of the majority of Hong Kong Chinese parents to accept such a change in policy since it was largely the aspirations of the middle class Chinese (known in the vernacular as the 'sandwich class') and not a desire by those in a position of power and influence to maintain the status quo by resisting attempts to accept Cantonese as the main medium of instruction[2]. Llewellyn (1982) noted that in order to reverse this public resistance "...one possibility is to embark on a long-term project of changing parents' and employers' attitudes towards Chinese as a teaching medium" (p29)[3]. It would seem that parental attitudes have changed little in the intervening time, since the Education Commission (1996) in citing evidence from Sze notes that 'the present Hong Kong situation is based on parental desire rather than educational planning...' (Part 2 p.84).

2. The use of native-speakers of English in the educational system.

The report crucially links the falling standards in English with the policy of localisation and comments that "The situation has deteriorated markedly, we are given to believe, since the effects of the policy of localization of teaching staff have begun to be visible. We consider the 'localisation of staffing' policy ought to be amended so that children in their first years of schooling might be exposed to native English speakers engaged as ancillary staff either on a contract basis or accepted as helpers (e.g. non-working spouses of British expatriates or other suitable English speakers..." (p27). These locally available native-speakers would for example be non-working spouses of British expatriates and could be employed as ancillary workers either on short contracts or on a part-time basis. They would function in the same way as classroom assistant and serve as a role model of native-speaker English for language learners. Thus for the first time there was an official linking of declining English standards to the policy of localisation, and the belief that this trend could be reversed, at least in part, by the employment of native English-speaking teachers.

Over the years however, the presence of NETs has been much stronger than it is today, and as Boyle (1997) points out "...native-speakers in the past were indeed assumed to be better teachers than local Chinese teachers of English. Now, that's no longer the case" (p.170).

II. THE EDUCATION COMMISSION

1. The First Education Commission Report (ECR1)

The Education Commission (1984) noted that Llewellyn (1982) was highly critical of the standard of English of both teachers and pupils and that the report advocated 'the employment of more native English speakers and fluent speakers of English in schools as English teachers' (Education Commission, 1984:32). ECR1 noted that some provision already existed for the recruitment and deployment of up to three NETS per school but most were not doing so partly because they did not want to become embroiled in the issue of providing housing for the expatriates (Boyle, 1997). The Education Commission felt that such problems could be dealt with and so the report went on to recommend that schools should be encouraged to employ "...locally available native English speakers with teaching qualifications to teach English" (Education Commission, 1984:39). In addition, ECR1 went on to recommend raising the quality of teaching English in schools by recruiting expatriate lecturers of English for the former Colleges of Education and the Institute of Language in Education (ILE). Thus, the relatively modest proposals put forward by Llewellyn (1982) regarding NETs, were transformed into something more substantial in ECR1. On a more general level, ECR1 notes (p.43) that 'all other things being equal, teaching and learning would be generally more effective if the medium of instruction were in the mother tongue.' Although there was some feedback from educationists on ECR1 (e.g. Johnson, 1985), discussions focused largely

around more general medium of instruction issues rather than areas specifically related to NETs.

2. The Second Education Commission Report (ECR2)

Between ECR1 in 1984 and the second Education Commission Report (ECR2) in 1986, very little progress was made. On the question of hiring expatriate lecturers for the former colleges, ECR2 merely notes that a recruitment exercise 'is in progress and a number of new recruits are expected to be in post by September 1986' (p.16). As regards the recruitment of qualified teachers whose mother tongue was English, ECR2 merely notes that 'We understand that the Government is developing proposals to recruit qualified English teachers...' (ECR2:16). As Boyle (1992) points out, the slowness of the government to act and the vagueness of some of the comments (e.g. concerning the recruitment of NETs) is surprising, since the government had hinted at more resources to help schools cope with expatriates' housing problems.

In fact, what transpired was that not everyone was as enthusiastic about the scheme as the government. Head teachers were beginning to voice concerns related to obvious practical difficulties both in the classroom and in the school generally. There was friction too in local schools due to a host of contractual, pedagogical and administrative questions (Johnson and Tang, 1993). Local teachers and principals also raised questions regarding the overall value of the NET Scheme in terms of predicted educational

outcomes. There is no evidence that any of these concerns were addressed by the Hong Kong government or any of its agencies involved in the scheme.

3. The Fourth Education Commission Report (ECR4)

> 'The Expatriate English Language Teacher Pilot Scheme was completed in July 1989. As a result of the final evaluation of the pilot scheme, it was decided that a permanent scheme should be introduced in September 1991.' (Education Commission, 1990:89).

Thus by the time ECR4 was released in 1990, it had already been decided that a permanent NET scheme should be set up in 1991. However, expatriate teachers only took up a minor part of ECR4, which dealt with the bigger issue of the medium of instruction (MOI) in H.K. schools that for some time had been a contentious issue. On the one hand, the business community felt that any move towards Chinese as a medium of instruction (CMI) would further erode Hong Kong's potential as an international financial and business centre and the English standards of future graduates would decline. Parents too, valued English as a medium of instruction (EMI) as it helped ensure the hegemony of the English language and their children's future: '...there is pressure for children to learn English and to learn *in* English, since this is seen by parents as offering the best prospect for their children's future. Many children, however, have difficulty with learning in English.' (ECR4, p.93). On the other hand however, educationalists had been arguing for more Chinese medium schools since most local students did not have a sufficiently high level of English language proficiency to benefit from EMI. In the end, the policy advocated by ECR4 was to strengthen CMI in primary schools, confirming the belief that

'the majority of students will learn more effectively through their mother tongue than through English' (p.95), while at the same time recognising 'the need for some English medium secondary education to be maintained and strengthened' (p.94). Consequently, it was recommended that while the top 30% of students could study through EMI, the majority 70% should receive their instruction through CMI and learn English only as a subject. Despite this urging by the H.K. government for a more radical change towards CMI, the majority of parents continued to resist any such changes, as they still wanted their children to attend EMI schools.

The issue of expatriate English language teachers was off the top of the educational agenda in the early 1990s as more general medium of instruction issues took precedence. With Hong Kong about to be handed back to China, localisation was in full swing and expatriates on lucrative overseas terms were considered privileged. In addition there was a large influx of immigrants from mainland China, adding to the burden being placed on teachers. Due to these factors, combined with the existing problems of housing, administrative problems at the school level and resentment from local teachers, in 1995 only 33 expatriate teachers were recruited into a total of 360 eligible government and aided schools.

3. The Sixth Education Commission Report (ECR6)

When ECR6 was published in 1996, evidence to date regarding the effectiveness of the various NET schemes in Hong Kong was at best mixed. Certainly, students taught

by NETs had not been disadvantaged and in some cases there was evidence of small gains in specific areas and at certain age and ability ranges. The evidence was far from being overwhelming. Thus, when the Draft Report (1995) stated that 'The Commission notes that the Expatriate English Language Teachers Scheme ... has been useful in improving the learning of English in secondary schools' (ECR6 Draft p.ix), writers in the field (e.g. Boyle 1997) expressed surprise. This Draft Report recommended that within two years (i.e. by 1998) all Hong Kong secondary schools should have two or more NETs who would be employed on local rather than expatriate terms. When ECR6 was published however, these recommendations had been toned down in what Boyle (1997) refers to as 'a note of realism' with ECR6 stating: 'we recommend that all government and aided secondary schools should be encouraged to engage more qualified English teachers who are native-speakers to fill graduate teaching posts on local terms of service' (p.53)[4]. It is interesting to note that while many of the recommendations in areas covered by ECR6 (e.g. language acquisition, medium of instruction policy, language planning) are supported by evidence from research (including comprehensive reference lists), the recommendation for the continuation and expansion of the NET scheme is not similarly supported by references to the literature or related research. Again, it would seem that this recommendation appeals more to common sense and the assumed logical connection between employing native-speaking English teachers and raising language standards than to any rational argument derived from the literature. Sweeting (1992) goes further, stating that '*Laissez Faire* is probably the most commonly used term to describe the general attitude of the Hong Kong government. Perhaps equally strong claims could be made for "benign indifference" or "enlightened inertia"' (p.72). Indeed, in referring to the

lack of development of language policy ECR4 concedes that '...what was possibly lacking was a coherent framework within which these [language improvement] measures could be conceived and implemented' (p.98). So far then, there seems to be little or no evidence from the literature on the local context that the deployment of NETs can be directly related to increased levels of language proficiency in Hong Kong schools.

Let us turn now to the more specific deployment of NETs in recent years in Hong Kong.

II. Native English-speaking Teachers in Hong Kong

1. The History of NETs in Hong Kong

The involvement of expatriate (i.e. native English-speaking) teachers in Hong Kong schools goes back at least as far as the period of British colonisation of the Territory. In the early days, the motivation had arguably more to do with spreading the gospel rather than British culture and the English language (Sweeting, 1990). This missionary work was not of course confined to Hong Kong – other noticeable 'targets' of Western churches were amongst others Macau, Malacca and later Japan. The recruitment of English native speakers from Britain is documented as far back as 1862

Up until 1982, recruitment of native speaking teachers was done on an ad hoc basis, but following the Llewellyn report (1982), whose recommendations were picked

up in ECR1, a more systematic and widespread of recruitment and deployment of native speaking teachers was adopted by ED.

As alluded to earlier, a number of problems began to arise with the NETS. Not least of these was the cost, estimated to be HK$50 million or more (Boyle 1992) if a large number of schools responded positively to the government's call. In addition to this however, other problems soon became apparent, among which were the following:

- an inadequate housing allowance;
- due to the exchange rate, a built-in erosion of Hong Kong expatriate teachers' salary;
- confusion over the number of Expatriate English Language Teachers (EELTS) required;
- the realisation by the expatriate teachers that it was not possible to bring about change;
- inevitable language problems at the school level;
- the lack of possibility or opportunity of EELTS to become involved in curricula issues.

There was much resentment by local teachers towards the EELTS scheme from the beginning. This was partly due to inadequate preparation from the government since the local teachers resented the implication that the EELTS were being brought in to show the locals 'how it should be done' and the expatriate teachers were under the impression that they were being brought in as 'agents of change.' Both parties were incorrect in their assumptions. Nevertheless, difficulties on the ground were very real, with expatriate

teachers receiving a higher salary than their local counterparts for doing the same job (because they received an additional accommodation allowance), a state of affairs obviously resented by local teachers. In addition, there were completely different methodological approaches, with expatriate teachers espousing communicative language teaching, often involving communicative language activities that were frequently seen as noisy games. Letters to local newspapers (e.g. Law, 1987) suggested less costly alternatives such as in-service training for local teachers and providing schools with better equipment and resources. Despite this, the majority of Hong Kong teachers were sympathetic to the principle of employing expatriate English language teachers.

The automatic superiority of the native speaker was thus at least being questioned, if not openly challenged by local teachers in Hong Kong. Implicit in the Education Department's EELTS scheme was the notion that expatriate English teachers were better than local teachers, but it was the latter that were more attuned to the needs and circumstances of the pupils. The local teachers knew the language, culture and the difficulties that their pupils were experiencing in learning English. The expatriate teachers on the other hand knew little or nothing of their pupils' language, culture or socio-economic circumstances; points summarized eloquently by Lung (1999):

> By giving recognition to a teacher's native-speaker ability, administrators had automatically marginalized local (nonnative-speaker) teachers. This marginalization demoralizes and diminishes the usefulness of local teachers and does a disservice both to the teachers and the students. For a start, local teachers have a clear understanding of the needs and backgrounds of the students, including cultural and linguistic factors. (p.8)

The 'problem' of course for Hong Kong was that there were simply not enough qualified teachers of English with a sufficiently high command of the language. The feeling was that the EETS would have met less resistance from local teachers and stood a higher chance of success if it had been introduced at least initially on a more modest scale and if the expatriates had been brought in as helpers working alongside locals and not as '...native-speaker specialists, expecting special treatment and special status' (Boyle, 1997:175). Missing too from the EETS was any systematic plan for in-service training of local teachers, as pointed out by Lung (1999): '...it would seem beneficial to investigate the possibility of a modest supplementary expatriate programme while focusing on the proper training of local teachers... [because]...Hong Kong's success in teaching English depends more on producing high-quality local teachers than on importing NETs' (p.8).

3. Declining Language Standards

> "For several years there has been a growing impression in Hong Kong that the standard of English is on the decline. Schoolteachers, university teachers and employers in the business sector all agree that their students and employees have a much poorer command of English than in earlier years." (Boyle 1997:163)

A constant theme in political, business and educational circles in recent years in Hong Kong has been the perceived falling standards of English language proficiency in all quarters, ranging from teachers to primary school pupils. Members of the Education Commission have expressed concern about the declining standards of English. Poon

(2000) notes that research carried out by the Education department has confirmed this. Compared with students in other Asian countries and in other parts of the world, the English standards of Hong Kong students are 'lagging behind' (Poon 2000:178). This, as Poon (op cit) points out, is the reason why language issues have been central to recent Education Commission reports. Johnson (1995) reports that a number of other studies 'reveal the inadequacy of many students' English language proficiency at various levels within the education system' (p.11). As far back as 1982, official government and quasi-government reports were referring for example to the "widespread concern with the alleged downward spiral in language competence in Hong Kong students"[5] (Llewellyn, 1982:25). The same report goes on to talk about "the present lamentable situation concerning the use of English as a medium of instruction..." (Llewellyn, 1982:26). Further, the visiting panel notes "Even in the upper secondary school we observed such low standards of English in both teachers and pupils that the essence of the lesson was largely lost." (p.27). The first Education Commission Report acknowledges the Llewellyn Report's findings, noting that it was 'critical of the standard of English of both teachers and pupils...' (Education Commission 1984:32). The Education Commission in its reports continues to refer to declining standards of proficiency in English. ECR6 notes that "Teachers, the business community and tertiary institutions perceive a decline in standards of proficiency in English and Chinese ... Whether standards have fallen and if so to what extent is not clear. It is clear that there is a widening gap between the level of proficiency demanded, particularly in English, and the level which the education system has been able to supply." (Education Commission, 1996: Annex 1B p.5).

Not all observers have agreed however with this public perception of falling standards. Towards the end of 1988, the Hong Kong government set up a working group within the Education Department (ED) to consider the issue of language in education. The working Group's report, printed in 1989 but not made available to the general public, is referred to in ECR4 and neatly sums up the counter argument: "On language standards, the Working Group concluded that English standards appeared to have been generally maintained but the fast increasing demand for competent users has led to a misperception that standards are falling" (Education Commission 1990:93). Thus it may be the case that standards are not in fact falling, simply that demand is increasing both in terms of student numbers entering for public examinations and going on to tertiary education. In the 1991/92 academic year for example, 31,486 students enrolled in undergraduate courses whilst in the 2000/01 academic year that figure had grown by some 52% to 47,880 (EMB, 2002:4). In addition, more complex language demands brought about by increased internationalisation and globalisation imply that the *quality* of language demands is rising possibly faster than the *quantity*. There is also some data suggesting that standards are rising. For example, the percentage of those claiming to understand English increased from 44% to 70% between 1983 and 1993 (ECR6:5). Bacon-Shone and Bolton (1998) review a wide range of empirical research on multilingualism in Hong Kong, looking in particular at government censuses. They find a big difference in people's perceptions in response to the question: "How well do you know English?" over the ten year period from 1983 to 1993. This change in people's perceptions is summarised in table one below.

Table 1. Summary results of survey on perceptions of how well people know English.

	1983	1993	increase/decrease
not at all	31.1%	17.4%	(- 14.3%)
a few sentences	23.5%	21.7%	(- 1.8%)
a little	36.2%	27.2%	(- 9.0%)
quite well	4.8%	26.6%	(+ 21.8%)
well	1.4%	3.3%	(+ 1.9%)
very well	0.4%	3.8%	(+ 3.4%)

(Bacon-Shone and Bolton, 1998)

Thus in the space of ten years those who say that they did not know English dropped dramatically from 31.1% to 17.4%. In contrast, those who reported that they know English very well rose from 0.4% to 3.8% - almost a ten-fold increase! We should of course caution this kind of survey since we do not, for example, know what 'quite well' means, nor do we know if respondents' views are consistent and reliable. Nevertheless, a consistent pattern does emerge: Hong Kong citizens report that their standards of English are rising. Further positive evidence is provided by Hirvela and Law (1991), whose study found that 53.6% of teachers surveyed reported that they felt 'comfortable using English with foreigners', while only a handful (5.8%) reported feeling 'uncomfortable'. By contrast, most teachers in this survey (81%) agreed that the standard of English among students had declined. Only 4% disagreed.

Thus, there is a *perception* that English (and indeed Cantonese) language standards have fallen sharply in recent years but there is conflicting evidence on this view. It is also difficult to make comparisons between language standards as revealed by public examination results within a 9-year universal education system and examination results from a pre-9 year compulsory education system. The fact is, we simply do not know for sure one way or the other, but perhaps this is not the real issue. What does matter is whether or not language standards are high *enough* for the demands of a post-modern, globalised and increasingly complex Hong Kong society. Expectations within the community have risen, especially for English although as ECR6 points out "it may be unrealistic to expect the education system to meet these expectations (particularly in English) in full" (p.5). Johnson (1994) explains the conflicting evidence thus:

> This high level of expectation largely explains the paradox whereby a spectacular increase in the numbers of Hong Kong people who feel able to understand English is simultaneously perceived as a failure on the part of the Education system to maintain standards" (p. 12).

In addition, as Johnson (1994) points out, provided Hong Kong maintains its present economic and social course, the level of demand for high language standards can be expected to rise and to merely maintain standards at their existing level would be inadequate.

In response to these widespread concerns, perceived or otherwise, about falling English language standards, the Hong Kong government implemented a number of

measures aimed at raising English language standards. One such measure was the introduction and expansion of the NET scheme in primary and secondary schools throughout the Territory (see Tung, 1997). Once the decision was made to expand the existing NET scheme to 700 teachers, there was evidently a need, sooner or later, to determine whether this expensive resource was effectively delivering the desired aim. In order to evaluate and monitor the effectiveness of the current NET scheme, SCOLAR (the Standing Committee on Language Education and Research) commissioned a project in June 1998, which was undertaken by a group of researchers at the Hong Kong Institute of Education (HKIEd) in November 1998. Thus the 'Monitoring and Evaluation of the Native-speaking English Teacher Scheme' (MENETS) was established, a project that was to last two years, finishing in November 2000. All of the instruments, methodologies and data referred to in this thesis have resulted from the MENETS project. The final report on the MENETS project referred to in this thesis (MENETS, 2001) has been submitted to SCOLAR and is in the process of being released to the public. This thesis focuses on one specific aspect of the MENETS project for which the author was responsible, namely the measurement of oral language proficiency of secondary students.

The next chapter considers the literature related to Native English-speaking teachers and contextualises the need, rationale and focus of this current study.

Footnotes

[1] This was a dispatch from the Governor of Hong Kong to the Secretary of State for the colonies, Earl of Carnarvon, 27 January 1878 (in CO 129/181 p.133ff).

[2] The policy making dilemma, as relevant today as it was in 1982 is succinctly put: "In Hong Kong where proficiency in English is necessary for economic and political reasons, there is a classic public policy dilemma: whether to jeopardize the educational progress of the majority (and perhaps endanger the culture itself) in order to guarantee a sufficient number of competent English speakers; or to value the whole group (and in so doing conserve the culture) but accept the loss in capacity to deal with the international environment and hence a possible decline in the economic prosperity" (Llewellyn 1982:30).

[3] It should be noted that this was not a recommendation that was taken up with any enthusiasm by the government at the time, and it is only in very recent years that some modest attempts have been made to do so.

[4] Shortly afterwards, in March 1996, the Education Commission's proposed funding for native-speakers was cut in half by the first local Chinese Financial Secretary's budget.

[5] Note the use of the word 'alleged'. The claim of falling standards is one that has often been made in recent years, but to date there is no empirical evidence to substantiate such claims. HKEA examination results suggest that more students than ever are passing English language public examinations, but as is often the case, once a charge is repeatedly made, it becomes a 'truth' in the eyes of the public.

CHAPTER TWO

REVIEW OF THE RELATED LITERATURE

This chapter discusses the literature related to the present study, which it is felt can be divided into the following areas:

- An understanding of the term 'native speaker';
- The linguistic and political issues surrounding the deployment of native speakers;
- The effectiveness of the native speaker in language learning;
- The history and effectiveness of the native-speaking English teacher in Hong Kong;

I. THE 'NATIVE SPEAKER'

1. Background

Arguably the most basic question underlying this study is 'What is a 'native speaker'? It is necessary to pose this question since this study aims at measuring oral English language proficiency gain that may be attributable to NETs, and secondly as Davies (1991) points out the native speaker has long been key to many aspects of linguistic study:

> "Applied linguistics makes constant appeal to the concept of the native speaker. This appeal is necessary because of the need applied linguistics has for models, norms and goals, whether the concern is with teaching or testing a first, second or foreign language, with the treatment of a language pathology, or with some other deliberate language use." (Davies 1991:1)

In linguistic studies and in language teaching theory and practice, the native speaker has for some time occupied a key position. Davies (ibid) is one of a number of writers to acknowledges their importance, noting that the native speaker is generally 'used by linguists' in two ways: firstly to represent an idealised model, and secondly to give an exemplar of such a model. Davies (1996) refers to what he calls the "bio-developmental definition" given by Bloomfield (1933) who considers the native speaker to be "The first language a human being learns to speak is his *native language*, he is a *native speaker* of this language' (ibid p. 43). In this sense, being a native speaker is a historic fact. Bloomfield (1933, 1984) points out that children learn by observing, participating and interacting with the people around them and this is especially true in the domain of language learning. In English language teaching (ELT), the native speaker has always been a kind of benchmark against which other speakers of English have been measured and in countries where English is a foreign language (EFL), as opposed to a second language (ESL) i.e. where it is not used socially, as for example in Hong Kong, it is often felt that a *native speaker* model is needed since this will increase the learners' ability to be understood internationally (Edge 1988). In modern linguistics, it is Noam Chomsky's concepts that lay at the heart of the discourse that has led to the belief in the superiority of the native speaker teacher. Canagarajah (1999) states that 'native speaker fallacy is anachronistic' (p.79) since there is, for example, no linguistic basis for the superiority of one dialect over the other. In addition, language learning is a creative, cognitive and social process that has its own trajectory not fully dependent on the teacher (much less

the teacher's accent). Canagarajah (ibid) also refers to and questions Chomsky's stand point on this issue:

> "Noam Chomsky's linguistic concepts lie at the heart of the discourse that promotes the superiority of the native speaker teacher. The Chomskyan notion that the native speaker is the authority on the language and that he or she is the ideal informant provides an understandable advantage to the native speaker in grammaticality judgments. However the very label native speaker is questionable" (p. 78).

Despite long-term acceptance of the term *native speaker*, there is some debate in linguistic circles as to the precise definition. A number of writers put forward different ideas (e.g. Stern, 1983; Crystal, 1985; Richards *et al.*, 1985). Whilst differences remain in exact definitions, it seems to be generally accepted that native speakers are "people who have a special control over the language, insider knowledge about 'their' language". Boyle (1997), quoting for Davies' (1991) study on the native speaker notes that native speakers are the models we appeal to for the 'truth' about the language, they know what the language is ('Yes, you can say that') and what the language isn't ('No, that's not English')" (p. 164). This is true not only in the linguistic domains of grammar, syntax and morphology but also in phonology. Some linguists such as Quirk (1985) have long stressed the importance of Received Pronunciation (RP)[1] and believes that 'the standard language is inevitably the prerogative of a rather special minority' (p.4), but others (e.g. Crystal, 1985) are less convinced, and some even reject the idea that native speakers can only be those who speak RP British English or Standard American English. Other varieties of English (e.g. Australian, Indian, Nigerian) inevitably give rise to varieties of native speakers of English. Indeed, there is now widespread acceptance of English as an

international language and of the notion that there is not just *an English* but in fact many *Englishes*. There is however a notion that 'when these are described as *the other tongue* or *nativized varieties,* the English of the ethnic Anglos is still there in the background as the central reference point' (Rampton 1990:97). Boyle (1997) concludes that provided the variety of English being used is intelligible internationally then there seems little basis for not accepting it as a 'standard variety of international English' (p.165) and those who speak that variety of English as their first language should really be considered 'native speakers of English'.

Jenkins (2000, 2002) distinguishes between English as an International Language (EIL) and English as a Foreign Language where in the former, English is being learnt for international communication and not necessarily for communication with native speakers (NS), whereas in the latter, the phonological model selects the NS accent that will have the 'widest currency' among the learner's target (i.e. NS) community. Jenkins (2000, 2002) points out that since non-native speakers (NNS) now outnumber native speakers (NS)[2] and this clearly has implications for ELT pedagogy such as the "...need to empirically establish phonological norms and classroom pronunciation models for English as an International Language" (Jenkins, 2002:83). Interaction in English typically involves no first language speakers whatsoever. A paradigm shift away from Regional Pronunciation (RP) and General American (GA) has arisen largely as a result of the diminishing number of RP speakers[3]. Within this context, the primary concern for NNS thus becomes their intelligibility and this questions the prevailing RP model both on the grounds of its appropriateness and its teachability. This paradigm shift inevitably

puts the attention on the L2 user rather than the NS and implies that an L2 user model is more appropriate for teaching and learning purposes. In proposing a greater acceptability of L1 accent, Jenkins (2002) cites Bordieu's (1997) reference to a 'legitimate discourse' in which the EIL speaker's phonology needs to be 'intelligible and acceptable' to the target international community (i.e. NNS).

Noting that almost without exception, language corpa is derived from NS sources (e.g. Collin's COBUILD, British National Corpus), Jenkins (2000, 2002) draws on her own data derived from classroom observations and puts forward a new proposal for EIL pronunciation teaching, with both core and non-core areas. Outside of the core areas, speakers would be unconstrained in their use of L1 features of pronunciation i.e. they would adopt and accept local phonological norms. One of the problems with such a new curriculum however, as Jenkins (2002) herself points out, is that the 'NS standard measure still reigns supreme' (p 85) and thus the prevailing ELT ideology still adheres to 'deficit linguistics' models. Unfortunately, research and intuition suggest that 'traditional English pronunciation teaching is destined to fail all but a small minority of L2 learners (p 86). In addition, such new proposals remain controversial since there is still the question of what the L2 learners him/her self wants to learn and while some L2 learners would react positively to new EIL pronunciation models, others would not[4].

Another key area of debate and controversy surrounds the issue of the 'ownership of English'. Who, for example, should be entitled to make key decisions on language standards for communication wholly between NNSs? Whereas traditionally this has been

the sole domain of the NS, more recent trends towards the 'democratisation of the English language' throw into question the role of the NS as the sole 'gatekeeper' of the [English] language. It is arguably no longer appropriate that the NS is the 'unquestionable authority of not just language ability but also of expertise in its teaching'. Jenkins (2002) also discusses '...the anachronistic terminology in use to describe the users and uses of English' (p 8), such as NS, NNS and EFL. Jenkins (ibid) cites Kramsch (1993) who argues that '...the notion for a generic native speaker has become so diversified that it has lost its meaning (p 49). Current notions do not, for example, recognise the varieties of English in countries such as Singapore and to label many speakers, who have learnt English as an L2 and have achieved bilingual status, as 'NNSs' is arguably inappropriate at best and offensive at worst. It might therefore be more appropriate and in tune with the times to substitute the term NS for 'monolingual English speaker' (MES).

Whilst Jenkins (2000, 2002) makes a valuable contribution to the NS-NNS debate, both in terms of a comprehensive study of the literature and in her proposal for a radical, new approach to EIL pronunciation teaching, her research is confined to the field of phonology. As we shall see in the discussion on communicative competence (Chapter 3, Section 1.1), whilst this is a vital area when considering the different domains that constitute oral (English) language proficiency, there are many others besides phonology. On the other hand, researchers should perhaps consider more fully the arguments put forward by Jenkins (2000, 2002) with regard to the perspectives and advantages that the NNS has over the NS[5] and integrate these more systematically in the research design. In

the case of this study, the marking criteria for the oral assessment (see Appendix I) might arguably take more into account the NNS perspective and consider inter-speaker and intra-speaker variability. Weighing against this consideration however, is the fact that for better or for worse, the prevailing phonological model in Hong Kong's secondary schools is the linguistic deficits model based largely on the NS and any criteria that deviate too far from this would consequently raise questions of validity.

Árva and Medgyes (2000) point out that none of the alternative phrases suggested for the term *native speaker* have stood the test of time and they agree with Paikeday (1985) that the native speaker is a useful term 'precisely because it can not be closely defined' (Árva and Medgyes 2000:356). Finally on the question of defining the native speaker, Davies (1995) adds that 'the native speaker is a fine myth: we need it as a model, a goal, almost an inspiration. But it is useless as a measure' (p.157).

Stern (1983) notes that the native speaker's output is a crucial point of reference in language teaching theory: "The native speaker's 'competence', 'proficiency', or 'knowledge of the language' is a necessary point of reference for the second language proficiency concept used in language teaching theory" (ibid p. 341). Stern's (1983) understanding of a *native speaker* derives from his theoretical construct of native-like proficiency, consisting of four key components, namely:

i) The intuitive mastery of the *forms* of the language;

ii) The intuitive mastery of the linguistic, cognitive, affective and sociocultural *meanings,* expressed by the language forms;

iii) The capacity to use the language with maximum attention to *communication* and minimum attention to form, and

iv) The *creativity* of language use. (Stern 1983:346)

The model of the native speaker is one that is firmly in place in both the field of language teaching and learning as well as SLA research. Quirk (1990) for example, believes that the world's English learning problems are best handled by native speakers (p.7). In recent years however, the role of the native teacher has come under question for a number of reasons, a key one being that the native speaker is hard to define. Medgyes (1992) notes that "From a sociolinguistic perspective...the native/non-native issue is controversial. It is equally debatable from a purely linguistic view. Efforts to define native competence or native-like proficiency have yielded inconclusive results" (ibid p. 341). Despite these controversies, Cook (1999) points out that there is some agreement that "the indisputable element in the definition of *native* speaker is that a person is a native speaker of the language learnt first; the other characteristics are incidental, describing how well an individual uses the language" and "L2 speakers can not be turned into native speakers without altering the core of meaning of *native speaker*" (p187). Edge (1988) refers to this definition as the '*accident of birth* sense' (p.154). Yet even here the boundaries are blurred. The example cited by Medgyes (1992) and by no means extreme example, is one of an Indian for whom English was the sole language of instruction and the language through which he has communicated professionally ever since. There are numerous cases and countries such as this, where English is a second language, that 'break the homogeneity of the native/non-native division' (p. 341).

Braine (1999) considers a number of different perspectives on the 'NS-NNS dichotomy' (native speaker – non-native speaker), noting that for non-native speakers of English it can be a highly personal issue. One such perspective is that of Thomas (1999) who explores what she claims are 'issues of credibility' the NNS teachers are forced to confront in their professional life. Thomas (ibid) points out that "...it is very disturbing that even some professionals involved in TESOL believe that being a native speaker of English is a necessary condition to teach English" (p. 6). In her study, Thomas (op cit) notes the discriminatory hiring practices and double standards manifested by learners as well as professional organisations. As a consequence, many NNS teachers feel that they have to work twice as hard as their NS colleagues and have to prove themselves as competent and effective L2 users before they are accepted as professionals.

If we accept then that the country of birth and a speaker's first language are no longer the *sole* factors in what constitutes a native-speaker, what then are the characteristics of a native speaker? Tay (1982) describes a native speaker as 'one who learns English in childhood and continues to use it as his dominant language and has reached a certain level of fluency'. Rampton (1990) prefers not to use the term 'native speaker' and suggests that the terms 'language expertise' and 'language loyalty' be used instead, noting that 'The concepts native speaker and mother tongue are often criticized, but they continue in circulation in the absence of alternatives' (p97). Crucially, Rampton (1990) argues that 'expertise is learned, not fixed or innate' (p.98) and the emphasis shifts from 'who you are' to 'what you know' (p.99). This assertion adds further weight to the

arguments of those who challenge the notion that the ideal teacher of English is a native speaker. Cook (1999) believes that this variable notion of *expertise* is related not so much to a defining characteristic of being a native speaker as it is to an issue of *quality.* Boyle (1997) observes that 'native-speakerness' is connected to three premises: the language one first learns, the amount we use the language and the proficiency level attained in that language. Davies' (1991) in-depth exploration into the identity and role of the native speaker in linguistics, together with Tay's (1982) and Rampton's (1990) perspectives are summarized by Boyle (1997) who believes that five characteristics are consistently cited and considered essential, namely:

1. inheritance/ birth/ early start;
2. expertise/ proficiency/ fluency;
3. continual use as dominant language;
4. loyalty/ allegiance/ affiliation;
5. confidence/ comfortable identification.

Of course, these characteristics can only constitute a working definition since they give rise to a number of yet more complex questions. For example, how early is an 'early start'? What is an 'acceptable' level of proficiency? In the second of the categories above, a Hong Kong teacher of English would in many cases be better at explaining a grammatical point than a native speaker[6], although the latter would instinctively 'know' if something was right. Further, a Hong Kong teacher of English may well lack the confidence in speaking English that a native speaker would have (even though his/her

language proficiency may be perfectly acceptable). In terms of intelligibility, it is likely that young language learners in Hong Kong would more easily understand a non native-speaking English teacher than they would a native speaker with a broad regional accent (north east England, to name but one example). A non-native speaker is also more aware of the linguistic problems being faced by the second language learner than a native speaker and is more likely to be bi-lingual or multilingual. Problems of language transfer are also better appreciated by the non-native speaking teacher since he/she has been through the same difficulties that the pupils are going through. Edge (1988) feels that two points need to be considered in using the native speaker as a linguistic model. Firstly, 'the best model for the students is not a foreigner speaking his or her native language, but the native teacher effectively communicating in a foreign language', and secondly 'the role of the foreign native speaker in such a situation is to partner and support the native teacher in his or her communication' (Edge 1988:155). The more appropriate model according to Edge (ibid) is the one of the local teacher seen to be enjoying using the English language in an exciting and creative way and using it to communicate with his/her students. The native speaker meanwhile is best deployed not in providing a model of correctness but in supporting the local teacher's attempt to communicate with the students. Finally, Edge (1988) feels that it is important for us to 'escape from the essentially nationalistic world-view of *native speaker/non-native speaker*...[and get]...involved in furthering an internationalist perspective in which users of English are simply more or less accomplished communicators' (p.157).

Cook (1999) takes issue with the existing assumptions regarding native speakers, in particular that they should be the only authority on 'correct' language use: 'Language professionals often take for granted that the only appropriate models of a language's use come from its native speakers' (p.185). Language teaching might benefit more from focusing less on the native speaker and more on the L2 user, and that 'The main benefits of recognizing that L2 users are speakers in their own right, however, will come from students' and teachers' having a positive image of L2 users rather than seeing them as failed native speakers' (ibid. p. 185).

Other writers go even further, challenging the very premise of the native speaker and the assumptions underlying it. Paikeday (1985), reminds us that the term *native speaker* is frequently used but rarely defined, and goes so far as to say 'The native speaker is dead!" and objects to the very term 'native speaker', preferring instead to use 'more or less proficient users of English'. In a similar vane, Edge (1988) suggests the use of 'more or less accomplished users of English', Rampton (1990) considers the concepts 'expert speakers' and 'affiliation' and Kachru (1985) suggests the use of the term 'English-using speech fellowships'. Thus there is controversy even surrounding the nomenclature, yet as Medgyes (1992) observes, 'their meanings tend to overlap and they are no less spurious than the concept of the native versus the non-native speaker' (p. 342).

2. Native speakers: beyond linguistics

> The belief that ELT is non-political serves to *disconnect culture from structure.* It assumes that educational concerns can be divorced from social, political and economic realities. It exonerates the experts who hold the belief from concerning themselves with these dimensions. (Holliday 1994:67)

Issues surrounding the native speaker in language teaching are not confined to linguistics. Rampton (1990) takes a sociolinguistic perspective and stresses the importance of considering the links between people and language in many different ways. He also reminds us that 'There are always ideological issues involved in discussions about who speaks what in education, and political interests often have a stake in maintaining the use of these concepts' (p.98). More importantly, Rampton (ibid) asserts that the problem with the concepts 'mother tongue' and 'native speaker' are that 'they mix up language as an instrument of communication with language as a symbol of social identification' (p.98). This is why educationalists should refer to speakers with a high level of proficiency as *expert* rather than *native* speakers. These preferred terms also mean that we should not assume that nationality and ethnicity are the same as language ability and language allegiance.

The role of the native speaker in the foreign language classroom is seen by Holliday (1994) in political terms and is an exercise of power and status. Holliday believes that a sociology and anthropology of the classroom is needed because issues of status, role and authority are involved. A macro view of the social context of teaching and learning is necessary since '...there is a growing realization ... that applied

linguistics ... is not sufficient to enable us to understand all we need to know about language teaching and learning' (p.14). Holliday (ibid) discusses 'The myth of expatriate success' (p.147) and cites research showing that expatriate teachers constantly fail to 'address the portion of the classroom culture that belongs to their students, preoccupied as they are with the technology of their perceptions of *their* lessons [and] much of the 'success' of these technologies is in effect mythical' (p.147).[7] The issue of the *native speaker* and *expatriate* then, clearly goes beyond purely linguistic considerations. Phillipson (1992b) takes on the 'native speaker fallacy' and challenges the notion that 'the ideal teacher is a native speaker, somebody with native speaker proficiency who can serve as a model for the pupils' (p.193). Why, he asks, should the native speaker be intrinsically better qualified than the non-native? After all, many of the native speaker's characteristics can be acquired by non-native speakers through better teacher training. Medgyes (1994) too does not accept the assumption that native speakers are necessarily better language users than non-native speakers, acknowledging that they are only 'potentially more accomplished users of English than non-native speakers' (p.12), and non-native speakers have an 'equal chance of success' (p. 103). Phillipson (ibid) cites several other writers in the field (e.g. Unesco, 1953; Britten, 1985; Kachru, 1986, 1991; Sridhar and Sridhar, 1986) to support his view that 'The untrained and unqualified native speaker is potentially a menace - apparently many of the products of the British education system recruited currently into EFL do not know much about their own language' (p.195). As Phillipson (1992b) rightly points out, if a non-native teacher has gone through the process of acquiring English as an L2 and has a keen awareness of the linguistic and cultural needs[8] of his/her learners, then the non-native teacher is in fact

better qualified than the native speaker. This viewpoint that the first language is not a problem but a *resource* to be drawn upon and exploited in SLA is supported by others (e.g. Kachru, 1994; Sridhar, 1994) who feel that the learner's first language can be used as a 'cognitive bridge' in learning the L2 and that 'periphery speakers can use their vernacular competence to relate English better to students from their own communities and help them integrate English more effectively into their existing linguistic repertoire (Canagarajah,1999:80). As some of the above arguments have already implied, effective language teaching is not just a question of linguistic competencies. Teaching is a multidisciplinary skill and the other attributes of the language teacher are clearly not restricted to the native speaker. Canagarajah (1999) poses the question: 'Is a native speaker necessarily a good teacher?' and in response, points out that 'Language teaching is an art, a science, and a skill that requires complex pedagogical preparation and practice. Therefore, not all speakers make good teachers of the first language' (p.80). Canagarajah (ibid) points to the work of Britten (1985) who argues that multilingual speakers may in fact have an even better grasp of the English due to a more advanced metalinguistic knowledge and language awareness and may consequently be more effective English language teachers than native speakers (Canagarajah (1999:80).

The difficulty with the layperson's perspectives on native speakers and language teaching and learning is that they originate from a time when teaching the culture and the language were one and the same, and the language learners were assumed to be wanting to connect the culture with the language from which it originates. However, times have moved rapidly on, particularly with the recent explosion of the Internet to an extent

where possibly 'universal norms for English teaching can ... no longer apply ... There is now a shift 'towards both linguistic and cultural emancipation, [signifying] the end of the era with the British and Americans as guardians of a monopolistic global norm' (Phillipson 1992b:198). This view of linguistic imperialism is supported by Pennycook (1990), who argues for a more 'critical applied linguistics' because language learning is inexorably linked to cultural and political dimensions and language teaching that does not acknowledge this is arguably more about assimilating learners than empowering them. Thus we need to apply 'a *principled postmodernism* [that] can help us move, in the first instance, towards a critical *applied linguistics*' (Pennycook ibid p.10). Boyle (1997) also employs the term 'linguistic imperialism' in referring to the original recruitment of NETs for schools in Hong Kong. Luk (2001) states that the first noted use of the term 'linguistic imperialism' was by Phillipson (1992) in referring to the political and cultural dominance by the English-speaking world through 'linguicism'[9].

There is a belief then by some writers that a form of 'linguistic hegemony' is at play in the deployment of the native English-speaking teachers. Lai (1999) for example, claims that the NET scheme in Hong Kong is a form of neocolonialism masquerading as a language improvement scheme. Lai (ibid) also sees a political and economic dimension to this issue, suggesting that the expansion of the NET scheme can be associated with a number of other measures that support the goal of western investors to maintain the political and financial status quo, particularly after the 1997 handover. One of the arguments supporting this viewpoint is that the effectiveness of the NET scheme in raising English language standards in schools in Hong Kong has not yet been

substantiated at least in terms of measuring language gain[10]. A study by Lo (1992) conducted a socio-historical study into the nature of language and analysed the (then) current situation of English use in Hong Kong. He concluded that English is perceived as a language traditionally representing authority and increasingly seen as an authoritarian imposition, particularly when the EFL teacher is an expatriate. Lo (ibid) also noted that many students in 'working class schools' feel that they are being forced to learn English that is not needed for their studies or for their work in lower socio-economic sectors of the local community.

It is this type of controversy surrounding EMI, NETs and the use of English more generally in Hong Kong schools, and in the view of the author the need to build more empirical evidence of measurable language gain that is attributable to NETs, that lies at the heart of the current study.

3. The native speaker and language learning

Although this study looks primarily at differences in *product* or outcome between those students taught by native speakers and those taught by non-native speakers, we should nevertheless consider differences in *process* – for example what actually happens in the classroom. If, as the current language and educational policy in Hong Kong asserts, native speakers have a positive impact on the development of the English language proficiency of students, such an impact must be attributable not simply to the fact that one group speaks English as the mother tongue, but also because in the

classroom different types of processes are taking place. In one investigation of these processes, Árva and Medgyes (2000) investigate the differences in teaching behaviour between native and non-native teachers and compare the stated behaviour with the actual (i.e. observed) behaviour of the two groups. As Árva and Medgyes (2000) point out, for a long time 'the mere existence of non-native speaking teachers of English as an entity different from native-speaking teachers was called into question' (p.355). The non-native speaker has often been held in disregard despite the fact that the evidence does not support such a view and throughout the world non-native speakers are by far the majority group. In addition, negative perceptions of the non-native speaker have not been seriously challenged since most contemporary classroom research has tended to focus on the learner rather than on the teacher Árva and Medgyes (ibid). Whilst most educationists would agree with Tajino and Tajino (2000) that the presence of two teachers, one native speaking and the other non-native speaking in the classroom at the same time would be ideal, for most contexts this is not realistic. Medgyes (1994) claims that native and non-native English-speaking teachers (which he refers to as NESTs and non-NESTs) are 'two different species with the following differences:

1. NESTs and non-NESTs differ in terms of their language proficiency;
2. they differ in terms of their teaching behaviour;
3. the discrepancy in language proficiency accounts for most of the differences found in their teaching behaviour;
4. they can be equally good teachers in their own terms.

Samimy and Brutt-Griffler (1999) investigate non-native graduate TESOL students perceptions of themselves as EFL teachers in terms of their linguistic competence, communicative competence, and teaching behaviours in relation to native speaker teachers. They found that the subjects saw themselves as different from their native speaker counterparts but the differences were perceived not only in linguistic terms (i.e. competencies), but also in their actual teaching. Non-native speaker teachers were considered to be more 'sensitive to the students' needs, efficient, and dependent on textbooks', while their native-speaker counterparts were considered to be 'informal, flexible, and confident' (p.141). Native speakers were not necessarily perceived as 'better' teachers (only 12% of the respondents considered them to be 'superior') but successful learning outcomes depended on a) learner factors (e.g. age, motivation), b) teacher factors (e.g. skills, training, experience), and c) contextual factors (e.g. amount of input, authenticity of materials). Thus differences are perceived and a native/non-native speaker construct is recognized, but this does not necessarily translate itself into notions of 'superior/inferior' teachers. In fact, Samimy and Brutt-Griffler (1999) report that their study echoes the findings of Reves and Medgyes (1994) in that the participants involved in the study did indeed report a difference between native and non-native speakers in both their linguistic and pedagogic behaviour. However, more interestingly, in response to the question 'who is the more successful?' the participants responded in descending order: both, non-native speaker, and then native speaker.

The dichotomy between the native speaker and the non-native speaker has thus given rise to a number of controversial issues in the field of applied linguistics. As we

have seen, many writers consider the question of native versus non-native speakers to be at best counter-productive, and at worst somewhat irrelevant. From the literature, a far more relevant question might be one of how well qualified the EFL teacher is and although the debate will continue, those in the field should perhaps 'seek or create opportunities to discuss issues related to professionals from diverse, multilingual contexts to raise their own consciousness and awareness' (Samimy and Brutt-Griffler 1999:143).

Of particular relevance here is the quality and quantity of the teacher's (English) language input and the role that this input has on the student's second language acquisition (SLA). SLA research and traditional language teaching methodologies based on the NS usually define language learners in terms of how they are *different* from native speakers. The tendency to relate the L2 learner to the native speaker frequently results in the use of a comparative discourse in which *success* and *failure* are used to compare the learner's language output to the native speaker. Many SLA research methods, such as error analysis, depend upon making comparisons between the SLA learner's language and that of the native speaker (Cook 1999). Tang's (1997) research found that non-native speakers also make this comparison and acknowledge that the native speaker is 'superior' in most aspects of language use. As a result, many non-native speaking teachers feel inadequate in their work and 'are ill at ease with using English accurately and appropriately' (Medgyes 1992:343). It is likely that the non-native speakers' progress is hampered not by a deficiency in the L2 *per se* but 'by a state of constant stress and insecurity caused by inadequate knowledge of the language they are paid to teach' (p.348). Despite the lack of much empirical evidence (at least in terms of enhancing

learners' language proficiency), many writers accept the usefulness of the native-speaker, given certain conditions. Tajino and Tajino (2000), for example, investigate how team teaching (one local teacher paired with one NET) can 'provide students with more opportunities to improve their communicative competence' (p.3). They conclude that this type of team teaching arrangement is most effective when 'all the participants, teachers as well as students, are encouraged to learn from one another by exchanging ideas or cultural values' (p.3). The 'ideal teacher' is not confined to the native or the non-native speaker category but both have their potentials in that:

"- the *ideal* NEST is the one who has achieved a high degree of proficiency in the learners' mother tongue;

- the *ideal* non-NEST is the one who has achieved near-native proficiency in English." (Medgyes 1992:349).

There are indeed differences between the two groups of teacher, and they should be acknowledged in order to help all teachers progress and reach their potentials.

Ellis (1986) notes that 'it is self-evident that SLA can take place only when the learner has access to L2 [second language] input. This input may be in the form of exposure in natural settings or formal instruction. It may be spoken or written. A central issue in SLA is what role the input plays' (p. 12). Ellis (1986) points out that the role of input in the process of SLA remains one of the most controversial issues in current research. In early theories, SLA was considered to be one of 'habit formation' gained through practice and reinforcement. By presenting language in suitable 'doses', continued practice would ensure automisation of lexical, structural and phonological forms. In this

respect, L2 learning was similar to any other type of learning, in which stimulus and response chains could be established, controlled and reinforced by the teacher. This Behaviourist view of language learning did not take into account however the fact that (language) learning is not purely an external phenomenon. These views were challenged by Chomsky's theories of the 1960s, which emphasised a more mentalist view of language learning in which the learner's 'language acquisition device' is activated by input, which merely serves as a trigger. More recent research into the cognitive domain of SLA supports this view (e.g. Skehan, 1998), although it is also beginning to show that mere exposure to L2 by itself is not enough (Ellis 1986).

A further debate in the field of SLA revolves around what precisely constitutes 'optimal input and whether it is a) carefully selected and graded by the teacher according to specific, predefined criteria, or b) merely a question of providing 'comprehensible input' as Krashen (1982) asserts. Sridhar and Sridhar (1986) show that the SLA paradigm which tends to dominate theory building in applied linguistics in much of the western world is not universally relevant .

There can be little doubt that whether a native-speaker or a non native-speaker are deployed in the classroom, the learning environment is likely to be affected considerably although the input of both native speakers and non-native speakers are equally relevant. We next need to consider the historical, political and educational factors that have resulted in the Hong Kong government investing such considerable resources in the current NET scheme.

II. THE HONG KONG CONTEXT

1. A Historical Perspective

As previously mentioned, the involvement of expatriate (i.e. native speaking) English teachers in Hong Kong schools goes back at least as far as the period of British colonisation of the Territory and the recruitment of English native speakers from Britain is documented as far back as 1862. In more recent years, the Expatriate English Language Teachers Pilot Scheme (EELTPS) was conducted from August 1987 – August 1989. It was initiated by the Education Department of the Hong Kong Government with the British Council contracted out to manage the scheme. The aim of the EELTPS was to "improve the standard of English in the participating schools[11]" which in turn was seen as "increasing pupils' motivation and interest, improving pupils' language learning (both process and product) and contributing to furthering the general aims of the schools." (British Council 1989, pi) It was intended to deploy a total of 84 expatriate teachers, working in pairs[12], and in the end 81 teachers were deployed, mainly in forms 1-3, in 41 different schools throughout Hong Kong. By January 1989 however, 22 of the recruited teachers had quit the scheme and at this stage, less than 50% of the schools involved expressed interest in continuing with expatriate English teachers when the pilot scheme came to an end.

2. EELT Interim Report

At the end of the first year of the EELTS pilot scheme, the British Council commissioned an interim report (1988) with the overall results inconclusive[13]. The report noted that "This was perhaps not surprising since, arguably, one would expect little in the way of results after an 8 – 9 month implementation period[14]" Although only an interim report, it is worth dwelling on some of the findings since they still hold considerable relevance even today. The report considered two key areas of the EELTS: 1) the relative achievements of pupils in English language classes, and 2) the effectiveness of EELTS in improving pupils' English language learning and their motivation. In this, one key area of difficulty and potential conflict to emerge was that of the teaching styles and the methodology of the expatriates who had two general approaches to their teaching:

1. Being derived from U.K. secondary education, it was characterised by three features:
   a) use of differentiation (e.g. in pupils' attitudes, motivations and competencies);
   b) use of classroom evaluation strategies (in order to operationalise the above);
   c) increased responsibility being placed on the learner.
2. A communicative approach to the teaching and learning of English.

As the British Council (1988) report concedes:

> Neither of these approaches is the norm in the majority of Hong Kong classrooms, where there operates by policy and practice a pattern of activity in which students of different ability levels may well follow the same course design and texts, prepare for the same examinations, and be treated as an essentially homogeneous group in the teaching of whom the teacher retains the central role. (Dr. Bowers, Annex 4 p.3)

As regards the communicative approach, Johnson (1988), also states that, 'the communicative approach ... is not easily compatible with the long-established teaching traditions in schools' (British Council, 1988 Annex 3, p.4).

In terms of the quantitative analysis carried out in the interim evaluation, the report concludes that 'students have not been disadvantaged while some have gained measurable benefit by exposure to EELST' (p.8). Paired sample t-tests were carried out in 28 experimental and 8 'non-scheme' schools on three groups of students, as follows:

Group 1: experimental schools, 4 classes, Forms 1-3, n = >4,000

Group 2: experimental schools, but non-EET classes, n = >2,000

Group 3: control schools, 2-3 classes, Forms 1-3, n = >2,000

Three language skills/areas were measured: HKAT[15] (i.e. general language proficiency), listening[16] and oral[17] skills. Of these three, only the HKAT results showed any significant differences between groups. The previous year's HKAT results were used as a pre-test and the 1987-88 results as the post-test. The only significant differences were found to be at Secondary 3, where some Expatriate English Teacher (EET) effect did emerge, as follows:

Table 2. Summary results of pre-test and post-test HKAT results EELTS Interim Report

pupil level	*group*	*mean*
High	Group 1	69.49
	Group 2	65.52**
	Group 3	65.99
Medium	Group 1	43.22
	Group 2	41.03**
	Group 3	38.30
Low	Group 1	19.46
	Group 2	16.70*
	Group 3	15.21

** $p = < 0.01$ * $p = < 0.05$

Two points need to be made regarding these results. Firstly, the analysis was restricted to paired samples t-tests and while observed significant differences may allow us to make some inferences about different groups, it is not a longitudinal analysis, so we must thus be cautious in making causal inferences. We can not say with certainty, for example, that differences in the scores above are necessarily attributable to an 'EET effect'. Secondly, as the report itself states, the HKAT results may not be related to any EET effect: '...at the outset it was thought possible that if an EET effect were to appear, it would be in the Listening and Oral Tests rather than the HKAT. It may be that what we are seeing is a pupil effect i.e. that for the children it is the HKAT rather than the Listening and Oral Tests that really matters' (British Council, 1988:8). This in my view is more likely to be the case than the other possible explanation put forward in the report: that students at this level (i.e. secondary 3) may be better equipped linguistically to

benefit from being taught by an EET because at this age and stage of English language proficiency, they might be better able to comprehend explanations and expositions from native speakers. (Annex 4, p.11)

Due in part to the bad press that the EELTS was receiving, and the somewhat ambiguous results of the interim report, it was never publicly released[18] but as Boyle (1997) notes, a copy did reach the press who were eager to highlight the more critical findings. Nevertheless, at the end of this pilot scheme, a second report was again commissioned (but again not released[19]).

3. EELT Final Report

> "Our main conclusion in this report therefore is that the EELTPS has improved English proficiency and helped change attitudes to English" (British Council 1989 p.55).

The EELT Final Report was if anything more positive than the 1988 preliminary report, three key findings of which were that:

1. the EELTs made 'a discernable contribution' to improving spoken English at the junior secondary level;
2. attitudes towards EELTs were generally positive and were matched 'by measurable improvement' in both general English and English listening comprehension;

3. expectations of quick improvement were disappointed, because as the report noted, raising levels of (English) language proficiency is not a short term proposition.

i) General English

In the EELTs study, pre and post test measurements were taken and whenever significant differences in favour of experimental groups were found, the adjusted mean scores were compared with those scores from students in the control groups. An 'EELT effect'[20] was found in HKAT scores where in level 2, a 7%-8% difference was observed between the mean scores of students taught by EELTs as compared to those students taught by local teachers, and in level 3 a mean difference of 'up to 19%' was reported (significant at the $p = < 0.05$ level).

ii) Listening

A significant EELT effect was observed in students' listening skills at level 3 where a 2% difference was observed between the mean scores of students taught by EELTs compared to those taught by local teachers (sig. at the $p = < 0.05$ level, two-year effect), as shown in the table below:

Table 3. Summary results of EELT two-year effect on students' attainment in listening

	Group	*mean*
S3	Group 1	39.16
	Group 2	38.52**
	Group 3	36.70

** p = < 0.01

iii) Oral

A significant EELT effect was observed in students' oral skills at S1 and S2 (two-year effect) (3-4%), as shown in the tables below:

Table 4. Summary results of EELT one-year effect on students' attainment in speaking

	Group	*mean*
S1	Group 1	23.20
	Group 2	22.46**
	Group 3	20.48

** p = < 0.01

Table 5. Summary results of EELT two-year effect on students' attainment in speaking

	Group	*mean*
S2	Group 1	23.58
	Group 2	20.21**
	Group 3	22.72

** p = < 0.01

Again, in the final report (British Council, 1989) only t-tests were used to determine if there were any significant differences between group and thus the reservations expressed above about their use in the interim report are also relevant here. In the Interim Report, there was some evidence of a EELTPS effect on Form three HKAT results for all ability bands and this effect was supported by a similar one-year effect in 1988-89. Whilst no listening or oral effects were claimed in the Interim Report, they now seem to appear after two years of the scheme in oral skills (S1 and S2) and listening (S3).

Throughout this time, concerns continued to be raised regarding perceived falling standards of English and as one of a series of measures designed to address this problem, the H.K. government decided to extend the EETS for an additional two years. This time, instead of recruitment being carried out centrally by the British Council, provision was now made for individual schools (in the case of aided schools) and ED (in the case of government schools) to do so. This change meant that the lines of reporting were via the individual school principal and ED and not via the British Council as was the case in the pilot scheme. Under this latest scheme, the Expatriate English Language Teacher's Modified Scheme (EELTMS) attracted 33 teachers in 1989-90 and 23 teachers in 1990-1991. In the first year of operation, 11 aided schools and 8 government schools joined the scheme, but in the following year, 5 aided schools withdrew.

What was the evidence that expatriate teachers were enhancing the language proficiency of students they were teaching under this scheme? In the final EELTMS report (Educational Research Establishment 1991), school principals and panel chairs[21]

admitted that as a result of the expatriate teachers, they had observed a remarkable improvement in students' speaking skills, especially in S5 students (although they reported only a 'slight increase' in their writing skills). A typical example of principals' and panel chairs' attitudes lies in the response to a questionnaire item asking if respondents were satisfied with the performance of the EELTS. The response was as follows: very poor = 4.3%; poor = 4.3%; neither = 26.1%; good = 47.8%; very good = 17.4% [mean = 3.7 on a Likert scale of 1 to 5]. Although on the surface the response to expatriates' effectiveness seems positive, two points arise from the methodology employed in this evaluation: Firstly, this type of questionnaire item does not provide firm evidence of effectiveness (which is not to say that this data is not useful). We can not, for example, infer from this data that the expatriate teachers were any more or less effective than their local counterparts, nor whether the English level of the students' they taught did increase and by how much. Secondly, the use of mean scores in interpreting the questionnaire seems problematic because in this example, if those respondents (26.1%) who chose 'neither' are counted as scoring 3 (and the mean score here was 3.7) this might be accrediting 'positive value' to a response when none was intended.

Thus up to this point evidence from the various schemes in Hong Kong that expatriate teachers were improving English language standards seems rather tenuous. It was surprising therefore when the Draft copy of ECR6 stated that 'The Commission notes that the Expatriate English Language Teachers Scheme (the scheme) has been useful in improving the learning of English in secondary schools' (ECR6 Draft, 1995 p.ix). Equally surprising was the recommendation of the Final Report of ERC6 (1996)

that the NET scheme not only be continued but expanded, and this without reference to any research literature to justify such a policy.

Finally, in the NET discussion, there is in my opinion a need to refer to the medium of instruction (MOI) issue, since there is evidence that this, rather than the placement of NETs in schools, is going to make the biggest difference in language attainment. Chan et al (1997) for example, investigated the relationship between the amount of English used in the classroom and student achievement and noted a small but significant relationship between the amount of English used by teachers in the classroom and students' achievement in different subject areas. Correlation values were small (ranging from .10 to .24) but Chan et al conclude that the more English that is used in the classroom, the greater the gain in student's proficiency. On the negative side however, the results showed that the amount of English then being used was not sufficient to achieve the Government's EMI objectives and that "...the present situation maximizes the costs of EMI in terms of students' educational and linguistic development while minimizing any benefits they can derive" (p. 80).

Footnotes

[1] RP used to be referred to as 'BBC English' or even 'the Queen's English'. In ELT and linguistics, the norm or standard English is now generally referred to as RP.

[2] Jenkins (2000) cites Crystal's (1988) estimation that there may be as many as 1,350 million L2 speakers compared to around 337 million L1 speakers.

[3] It is estimated that fewer than 3% of the British population actually speak RP, with regional accents or 'modified RP' being the norm among NSs.

[4] Jenkins (2002) also points out a further controversy in that there is still to-date no academic course entitled 'English as an International language'.

[5] Jenkins (2000) refers to three advantages that the NNS has in particular, namely: phonological and phonetic knowledge systems, the intelligibility criterion and classroom pronunciation models (p 221).

[6] The point here being that often Hong Kong students of English will have a sound theoretical background to the (English) language since it has until recently been taught from a structural/grammatical approach. Many NS might know that something is 'right' or 'wrong' but they will often not know why.

[7] Holliday (1994) also provides evidence of success with expatriate teachers, which proved that 'the intrusion of expatriate lecturers did not have to be counterproductive' (p.155). These cases were ones in which the teacher had successfully integrated into 'the classroom culture' i.e. they have built into their repertoires 'opportunities for observing and learning about the culture of their students' (p.159).

[8] Holliday (1994) notes that the problem lies not only with teachers from different national cultures to their learners, but 'teachers are inevitably of a different culture from their students no matter what their nationalities' (p.159).

[9] Luk (2001) quotes Phillipson (1992) in defining 'linguicism' as referring to "...ideologies, structures and practices ... used to legitimate, effectuate, and reproduce an unequal division of power and resources (both material and immaterial) between groups (p.47).

[10] This is not to deny that gains have been made in affective factors such as motivation (e.g. MENETS, 2001).

[11] According to the contract between the Hong Kong Education department and the British Council (British Council 1989)

[12] It was felt that deployment in pairs would reduce feelings of isolation (except in schools where expatriates already existed on the staff).

[13] Rex King, Head of HKEA commented that "From my own experience I would expect one in every three overseas teachers to disappoint in terms of relating to Hong Kong and to the particular challenges they will face in Hong Kong schools...In addition, most teachers experiencing this kind of challenge for the first time would readily admit that it was only in their second year that they were able to make a worthwhile contribution." *South China Morning Post,* 24 February 1987.

[14] This was also found to be the case in the current study.

[15] Hong Kong Attainment Test – a standardized test of English used for norming purposes and administered to all students in the first three years of secondary school in Hong Kong. The English language component focuses mainly on usage and writing but also has a small listening component.

[16] The listening test devised for the EELT evaluation was specially devised by the Hong Kong Examinations Authority (HKEA) and was considered 'highly reliable'. There were a total of 60-70 items which were considered to be a better indicator of listening performance than the short listening section in the HKAT (see British Council 1989)

[17] Again, the oral test was specifically devised for the EELT evaluation and consisted of 37 items which tested lexis, accuracy and limited appropriacy.

[18] This further raised the suspicions of teaching bodies such as the Professional Teachers Union who one suspects would have grasped at the lack of evidence produced by this report.

[19] "This was even more extraordinary, since whereas the first, interim report had been, in the words of one British Council officer, 'neutral to slightly positive', this second report was 'very clearly positive'. (Boyle 1997:175)

[20] In the EELT final report (British Council, 1989), the reported percentage difference is that of comparative *mean differences* and is not to be confused with *effect size.*

[21] In the Hong Kong system, panel chairs function in much the same way and perform similar duties as Department Heads in countries such as the UK.

CHAPTER THREE

METHODOLOGY, INSTRUMENT DESIGN AND PROCEDURES

Having conducted a review of the related literature, and as a precursor to designing a test instrument, it is necessary first of all to explore the theoretical basis and the constructs underlying the notion of 'communicative competence'. This is carried out in the first section of this chapter, below. In English language Teaching (ELT), having a thorough understanding of what it is to be 'good at' English (i.e. to have a high level of communicative competence) is essential, since it is these constructs that inform the current literature and research on (English) language testing in general and the testing of oral language proficiency in particular. The second section of this chapter therefore explains how the theoretical constructs of language proficiency were used to design and develop the oral assessment instrument employed in this study according to our up-to-date knowledge and understanding of language testing theory. Then, having shown how the test instrument was constructed, this chapter will go on to present details of how this was piloted and subsequently revised in order to arrive at the final instrument employed in this research. Specifications of the assessment instrument itself will be given, including the processes, test stages and the criteria that were used to produce the data for analysis from the two administrations. Finally, this chapter will consider briefly the reliability of the test instrument by giving details of a post hoc reliability study conducted on a small number of students and trained raters. From this small-scale study of a specified sample, we may then reasonably make inferences about the reliability of the instrument as a whole.

I. CONSTRUCTS OF LANGUAGE PROFICIENCY ASSESSMENT

1. Communicative Competence

This section considers the development of the theoretical framework underlying the issues of oral proficiency, language testing and communicative competence. The rationale for the oral assessment instrument developed for use in this study is based upon an understanding and operationalisation of these theoretical constructs.

Since the 1960's, linguists' understandings of language were largely defined in terms of grammatical knowledge. Chomsky (1965) challenged this structuralist view of language arguing that proficiency can not simply consist of a set repertoire of phrases and sentences and proposed a more dynamic picture of language proficiency. Chomsky (ibid) posited a dichotomy between performance (i.e. ability to *use* language) and competence (i.e. the intuitive *knowledge* that an 'ideal' speaker-listener has of the linguistic system) and this dichotomy was for some time the prevailing model underlying language teaching and testing. Olivares (1998) reports however that this model came under criticism (e.g. Paulston, 1990), because it was unable to explain language outside the very restricted confines of an ideal speaker-listener in a homogeneous society i.e. a native speaker in a monolinguistic context. Hymes (1971) challenged Chomsky's ideas, claiming that they were too narrow since they did not include a social dimension to the language being described. Hymes (ibid) introduced the now widely accepted term 'communicative competence' which entails the ability to not only use the language but also knowledge about the language itself which is

necessary since "...there are rules of use without which the rules of grammar would be useless" (p.278), and, "...put otherwise, there is a behaviour, and, underlying it there are several systems of rules reflected in the judgements and abilities of those whose messages the behaviour manifests" (p.281). Ellis (1985) sums up the shift in emphasis succinctly: "The user not only knows what is correct, but also what is appropriate for each context of use" (p.78). Attention thus shifted from the purely *internal mental processes* of the individual to the *interaction* of the individual with his/her interlocutor in the process of *negotiating meaning*. One of the key concepts underlying the move towards communicative language teaching was the notion that language is above all about communication and this requires the negotiation of meaning. Thus while Chomsky overlooked stylistic variability, for Hymes (1972) and others espousing a more dynamic model, stylistic variability was instead an integral component of a speaker's communicative competence. This notion is now widely accepted and needs to be considered in developing communicative tests of language proficiency although there still remains some debate on what 'communicative competence' means.

Canale and Swain (1980a, 1980b, 1981) developed a theoretical model of communicative competence that includes four different competencies:

- grammatical (vocabulary, word formation, phonology and spelling);
- discourse (knowledge of textual conventions);
- sociolinguistic (rules of language use);
- strategic (verbal and non-verbal communication to compensate for shortcomings in the learner's L2 and communication breakdowns).

Canale (1983) extends and refines this definition of strategic competence further so that it includes not only the compensatory function advocated by Canale and Swain (1980a, 1980b, 1981), but in addition includes the enhancement characteristic of production strategies. Thus Canale (1983) defines his more refined construct of strategic competence as follows:

> "...[the] mastery of verbal and nonverbal strategies (a) to compensate for breakdowns in communication due to insufficient competence or to performance limitations and (b) to enhance the rhetorical effect of utterances" (p. 339).

Interestingly however, the proposed models above have not been *empirically* validated, with the result that new models have been developed that are more applicable to language testing. Bachman (1990) for example, proposes a theoretical model of communicative language ability (CLA) for language testing purposes. This model was further developed by Bachman and Palmer (1996) who also drew upon Canale and Swain's (1980a, 1980b, 1981) and Canale's (1983) models of communicative competence to develop a theoretical model of *language ability*, specifically for use in language test development. Within this model, language ability is specified within an interactional framework, involving two sets of parameters affecting language use and language test performance, namely: 1) the characteristics of the task/test situation and, 2) the characteristics of the individual test taker. Bachman and Palmer's (1996) model proposes that language tests should emphasize the communicative ability of the test takers within meaningful contexts in which they are likely to need to use the language. This proposition in turn raises two further conditions, namely that of *authenticity* (i.e. the need for the test to match the language use situation), and that of *interactiveness* (i.e. the need to ensure that the test-taker's

language ability is actively and directly engaged in accomplishing the test task). Bachman and Palmer (op cit) also specify five characteristics of the test task, which must be considered in test design: the setting/context, the test rubric, the input, the expected response and the relationship between the input and the response. Thus both the test takers and the test situation influence language use and test performance and must therefore be considered in test design so that they can "...facilitate rather than impede the test taker's performance" (p.61).

2. Testing Oral Proficiency

When testing oral language proficiency, a key question facing designers and administrators is 'what is language proficiency'? The question is not easily answered and there remains a wide range of opinions by different writers in the field. Clark (1975), for example, states that it is "...the student's ability to communicate accurately and effectively in real-life language use contexts, especially in the face-to-face conversations typical of the great majority of real world speech activities" (p. 23). Van Lier (1989) says that "...oral proficiency consists of those aspects of communicative competence that are displayed and rated in oral proficiency interviews" (p. 493). Olivares (1998) considers this question at length and cites the views of a number of writers such as Lantolf and Frawley (1988) who assert that "...we are no closer to understanding the concept [of language proficiency] today than we were 20 years ago" (p. 23). Olivares (op cit) also reports Spolsky's (1985) more cryptic comment that "...it has turned out to be simpler to think up new tests and testing techniques than to explain precisely what it is that they are measuring" (p. 23). Although there is no consensus on the precise definition, there remains some general

agreement that oral proficiency involves not only grammatical aspects but also an ability to *use* language appropriately and effectively in different contexts. It is also generally agreed that oral language proficiency includes notions of effective language use to achieve a range of communicative functions.

An understanding of these issues surrounding the nature of communicative competence, language proficiency and language testing is central to this thesis. It is upon these principles that the oral assessment instrument used in this study was designed and administered and it is felt that these constructs form a key argument for the construct validity of the assessment. Let us now turn to the specific instrument design and test procedure.

II. INSTRUMENT DESIGN AND TEST PROCEDURE

Given the emphasis in the job description of the NETs (Educational Department, 1997), it was hypothesised that for speaking there were likely to be differences between the performance of students taught by NETs and those taught by local teachers. Such differences in assessment outcomes of students' oral English language ability were likely to manifest themselves in terms of ease of delivery, fluency, obtrusiveness of pronunciation errors, effectiveness of communicative competence (i.e. strategies used) and grammatical accuracy as measured on the instrument employed in this study.

1. Instrument Design

The design and development of the secondary oral assessment instrument and procedures drew on a number of sources in additional to the theoretical constructs discussed in the above section. Firstly, the bands of performance for Key Stage 3 (Secondary 1-3) and 4 (Secondary 4-6) as developed by the Curriculum Development Institute (CDI) were referred to in order to predict target language for elicitation. In addition, the new Syllabus for Secondary Schools (1999) provided further input on linguistic content as well as subject/topic areas. A number of contemporary course text books such as 'New target English', 'English – A Modern Course' and 'Longman English Express' were also referred to in order to make use of the content, topic and task types that students in secondary school in Hong Kong are familiar with.

An important principle of the instrument design was that during the oral assessment, students be used to performing in the type of task that they commonly undertake in their classes. Thus the 'find the difference' task that has for some time been commonly used in TEFL textbooks (e.g. Ur 1981, Watcyn-Jones 1997) and is now in use in contemporary Hong Kong secondary school text books (e.g. 'New target English') was felt to be particularly appropriate. This was borne out by the feedback from oral assessors and the ease with which the tasks appeared to lend themselves to the assessment procedure and is important in addressing other key areas of validity. In order to help alleviate test bias, it was also important that the subject or content areas used in the assessment were ones that we could safely assume students would be familiar with. Thus the picture topics 'at the park', 'at home', 'in the

classroom' and 'the picnic' were found to be appropriate and suitable and ones that students are not unfamiliar with either in their daily lives or indeed in the English language classroom. These were considered to be important issues to address and build into the assessment design since they help satisfy concerns regarding the face validity and content validity of the assessment tasks and instruments.

The following section describes the development and piloting of the oral assessment instrument.

2. Development and Piloting

Three task types were tried out for the oral assessment for secondary one and three. They were all designed for administration to students in pairs, seated facing each other. First, a *Communication Task* was designed in which pupils are given two similar pictures. Based on the differences, pupils are asked to describe them or to find a mutually agreeable time for an appointment. Pupils do not see each other's pictures and are encouraged to contribute equally to the task. Secondly, a *Role Play* task was designed in which pupils were given a description of a situation and expected to take on roles and interact with each other in improvising a conversation. Thirdly, an *Open Discussion* task was designed in which the examiner was encouraged to develop a more natural, open and spontaneous dialogue and conversation with the pupils as an extension of the two previous tasks. Sample questions for the elicitation of more spontaneous speech are provided for assessors. Class teachers were asked to pair students according to their own criteria but were requested to ensure that students were familiar with working with their partner.

Pilot oral assessments were carried out at in a secondary NET school. The purposes were three-fold: to pilot the assessment instrument that had been developed, to obtain demonstration sample video and audio tapes for the training of assessors and finally to gain some baseline data for students' language proficiency.

Table 6. Student sample of pilot oral assessment

School	Class	High	Medium	Low	Total
Secondary	F. 1	5	6	5	16 (8pairs)
	F. 3	5	6	5	16 (8pairs)

Note 'High', 'medium' and 'low' in this study refer to relative ability of students.

3. Results of Instrument Piloting

As a result of the piloting, major revisions had to be made to the materials used due to the inappropriate level of difficulty of two of the communication tasks and all four of the role-play tasks. The feedback from assessors was that the inherent nature of the tasks was too complex for the students, there was an 'information overload' in the texts required to complete the tasks and that the materials did not effectively elicit the best possible sample of language from the students involved. In the revised secondary materials, the role-play section was replaced with a picture description task and two communication tasks. In addition, two more samples of communication tasks were added with visual illustrations and a graphic artist was also considered necessary in order to make the materials look more professional and to motivate the students to produce better language samples.

The final Oral Assessment instrument that evolved from the piloting process and an analysis of the feedback was used at both junior secondary (Form one) and at senior secondary (Form three, Form four). It consisted of three main stages: communication task, picture discussion and open discussion. Students were again interviewed in pairs that were determined by the regular class teacher. Teachers were asked to ensure that the pairs were 'compatible' and of similar ability to avoid one student becoming too dominant. Students were given five minutes to prepare for the interview beforehand during which time they were given an information sheet, in Chinese, explaining the interview procedure and a mark sheet on which they completed their personal details (name, class, teacher, etc). Assessors were advised to employ student helpers (peers of those being interviewed) to act as timekeepers and ushers and to distribute the documents to the candidates beforehand. After a 5-minute preparation period, candidates entered the interview room in pairs, where the assessment was conducted.

According to a standardized procedure, assessors greeted the candidates naturally, trying to put them at their ease. Candidates were asked if they had had sufficient time to prepare and if they understood the interview procedure. If they didn't understand the procedure, the assessor explained this to them. Up to this point, candidates and assessors could use Cantonese, but beyond this point the assessors were instructed to speak only in English[1]. The oral assessment procedure is explained in detail in the following section.

III. ASSESSMENT PROCEDURE

1. Oral Assessment Instrument

Both students being interviewed were asked standardised warm-up questions e.g. "Hello, how are you?" "What's your name?" "How old are you?" "Which class/grade are you in?" The purpose of the warm-up stage was to ease the candidates into speaking English in a relaxed, informal manner. This 'pre-stage' of the interview was not assessed and following a suitable period was followed by the formal assessment which consisted of three stages, as follows:

i) Stage 1: Communication Task

Candidates were seated facing each other and both were given a picture, which was similar but not identical. They were not allowed to see their partner's picture but were instructed to find up to 10 differences through asking their partner questions, describing and discussing the pictures[2]. Assessors encouraged the candidates to contribute equally to the task and if one pupil was dominating they encouraged the quieter of the two to participate more. At the end of this pair work communication task, both candidates were asked to summarise a few of the differences that they had found during their conversation. All the interviews were recorded on audio tape, and a smaller number were filmed on video tape for subsequent moderation and second stage analysis[3].

ii) Stage 2: Picture Discussion

In this second stage of the interview, the candidates were given another picture which this time they had to describe and discuss. The aim of this was for the assessors to elicit as rich a sample of language as possible from the candidates by asking them to talk about the pictures. If they were unable to do so or if little language was elicited, specific questions were asked from a standardised list.

iii) Open discussion

At the end of the second stage, assessors tried to develop a more natural, open conversation with the candidates in a kind of 'extension' of the picture discussion but which was more personalised and related to the candidates own lives and experiences. Again, a set of standardised sample questions designed to elicit more spontaneous speech was available in the assessors' test file packs.

During the oral assessment standardisation workshops (see below) the assessors were trained to put the candidates at their ease and to elicit as rich a sample as possible of spoken English language by progressing from the general to the specific. To achieve this, a standardised questioning technique[4] was employed which involved:

- starting with open questions. If this was unsuccessful,
- asking -wh questions. If this was unsuccessful,
- progressing to increasingly more specific questions. Then,
- asking choice questions, and finally,

- asking yes/no questions.

The interviews were designed to last for 12-15 minutes for Form one students and for fifteen to twenty minutes for Form three and Form four students. At the end of each interview, when the students had left the room assessors completed the mark sheets according to the assessment criteria in the test file packs.

2. Assessment Criteria

The assessment criteria that were employed for the secondary oral assessment were derived from a number of sources. The theoretical constructs of oral proficiency and oral assessment were investigated with reference to Bachman (1990, 1996), Hughes (1989), Weir (1990, 1993).

As a result of the assessment piloting and feedback from assessors during the training workshops, some modifications were subsequently made to The Centre for Applied Linguistics rating scale for CAL oral proficiency (CAL, 1998). This revised scale was used as the basis for the final descriptive criteria used throughout the secondary oral assessment. These criteria consisted of five elements: Comprehension/ Communication, Fluency/Productivity, Vocabulary, Grammar and Pronunciation. Within each of these elements, there were 6 levels numbered from one (low level) to six (high level), resulting in possible overall scores ranging from thirty to six (although in a small number of cases assessors awarded zero in some categories).

IV. METHOD AND NATURE OF SAMPLING

1. Sampling

Multi-stage stratified random sampling was used in the MENETS project to develop a representative sample of schools. To this end, questionnaires were sent out to all schools employing NETs in February 1999 asking them to provide information about the deployment of the NET in their schools. The results of this questionnaire indicated two principle types of NET deployment, namely 'oral only' and whole/split class. In the former, the NET was usually deployed to teach at least one lesson per week (or per cycle) to all classes in the school. In the latter, the NET was assigned up to four whole classes or parts of classes and asked to teach *all* the English lessons to those classes.

A second variable affecting the drawing up of a representative sample was the starting date of the NETs. In the 1998-1999 school year, NETs were recruited in two batches: September 1999 and January/February 1999.

Based on the above information, it was possible to identify the following four strata of NETs:

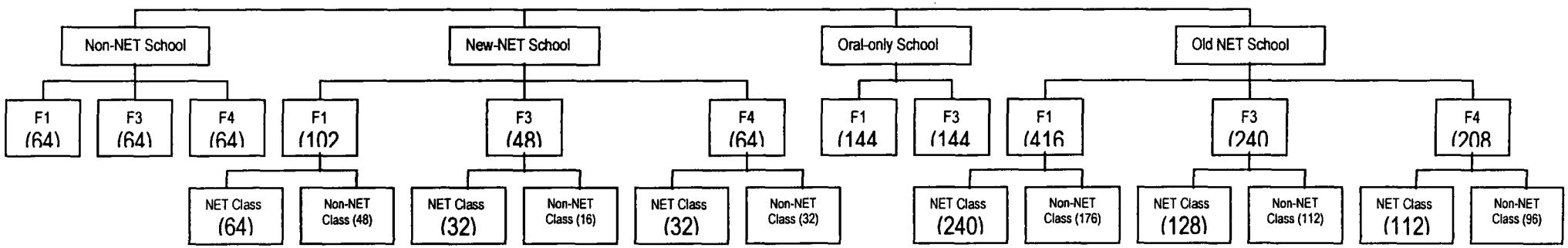
1. non-NET schools
2. new-NET schools (schools employing NETs post January 1999)
3. oral-only NET schools (in which NETs taught oral classes only)
4. old NET schools (schools employing NETs pre January 1999)

The identification of schools that were suitable for this research had to take the above four strata into account, but as pointed out by MENETS (1999), although the MENETS team were attempting to establish an ideal multi-stage stratified random sample this was not ultimately possible. Many schools simply refused to co-operate and alternative, replacement schools had to be identified. Finally however, after much negotiation forty nine schools agreed to become involved in the project, of which only four schools were non-NET schools[5]. However, since the purpose of the non-NET schools was in any case to provide control groups for the 'oral only' NET classes, this was not felt to be critical. It was agreed that this data could be collected from students not taught by NETs in the NET schools.

Having identified the schools that would be suited to the research design, it was then necessary to select sample students for assessment from these schools. In order to achieve this, students taught by NETs in Form one, Form three and Form four were first selected (NET classes), then students taught by local teachers of a similar ability level were selected for comparison. Finally, it was possible to build a complete sampling profile for the oral assessments which is illustrated in Table 7 below.

Table 7. School and Student Sampling for Assessments

Oral Assessments



Selection Criteria: *16 students were sampled from each class according to student English ability as follows:*
High : Medium : Low = 5 : 6 : 5

Total number of secondary students involved in the oral assessments: 1558

Note: Figures in bracket represent the number of students sampled

(MENETS, 1999:18)

2. Timing

The MENETS project was commissioned to be conducted over a two-year period, from September 1998 to August 2000. It was envisaged that as part of this project, the Time one (i.e. pre test) oral assessments would be conducted shortly after the beginning of the 1998 – 1999 school year, with the Time two (i.e. post test) being conducted near the end of the 1999 – 2000 school year. It was felt that a period of almost two years was necessary in order to accurately measure any gain in oral English language proficiency. However, these target dates proved to be over-ambitious.

Firstly, it was necessary to go through a number of stages before conducting the pre tests, namely: theoretical/construct design, initial instrument design, piloting, revision of test design, oral assessor training and finally the conducting of the actual assessments. These stages can prove to be problematic, especially when large numbers of schools, trained assessors and students are involved. In short, the pre tests were not in fact conducted until February-March 1999.

Secondly, it was not possible to conduct the post test at the very end of the 1999-2000 school year, since at that time most schools are involved in public examinations, internal examinations and a number of extra curricular activities. In addition, those form four students who are going to leave school usually do so at Easter time and do not attend the final, summer term. For these reasons it was necessary to conduct the post tests in

May 2000. Thus while a two-year time line had been planned between the pre and the post tests, in reality this period turned out to be little more than one year.

3. Loss of Data

As we can see from table 7 above, in the original research design it had been intended to administer the assessment to 1,558 students in both the pre and the post tests. In fact however, there was considerable loss of data. Due to teacher and student absentees, unexpected extra curricular activities and administrative difficulties such as teachers and/or schools withdrawing their support from the project at the last minute for a number of reasons (in many cases very justifiable), the pre test was finally administered to a total of 1,424 students.

It was of course intended to administer the post test to the same 1,424 students but in fact this number decreased to only 916. There were a number of reasons for this loss of data, including again teachers and/or schools being unable or unwilling to conduct the second administration of the assessment, schools having to focus attention and resources on internal and public examinations and what schools perceived to be an unnecessary drain on their time and resources. One difficulty faced by the MENETS project was that the team was unable to exert sufficient pressure on the schools/teachers to comply. We must also bear in mind the fact that this oral assessment was only one small part in a larger project which involved the large scale administration of questionnaires, direct interviews with teachers student and principals, school case studies and classroom observations. Whilst most schools were in general helpful and cooperative in their

support of the MENETS project, some clearly felt that they could not justify the resource allocation required for the administration of the second administration of the oral assessment.

As a result, many of the tables contained in chapter 4 refer to 'missing data' and detail the precise numbers missing between the pre and the post tests. It is the lost data described above to which this may be attributed. Finally, on the question of missing data in the subsequent chapters on Rasch scale modeling (Chapter 5) and Multi level modeling (Chapter 7) complete data sets are required for these types of statistical analysis. That is to say, not only is a pre and a post test score required for each individual student, but also the *complete list* of all other variables such as student ability, teacher, school level, etc (see Chapter 4) are required.

The next section will consider the issues related to test reliability that inevitably need to be considered in a study such as this one.

V. RELIABILITY

1. Internal Consistency

Cronbach's Alpha values for both administrations of the two cohorts were calculated with results as shown in table 8 below.

Table 8. Internal Consistency of Secondary Oral Assessment

	1st Admin	2nd Admin
Secondary Oral Assessment	$\alpha = 0.97$	$\alpha = 0.96$
	n = 1426	n = 928

NB These Alpha score could not be raised by deleting any of the items

The above Alpha figures are sufficiently high to indicate that the items within the test are tending to measure the same trait/construct. However, given these unusually high figures, there is some question as to whether or not assessors were able to discriminate effectively between levels using the given descriptive criteria in the five categories, or whether in fact they 'fixed' in their own minds an overall total score and then proceeded to mark all of the separate categories. Nevertheless, the high Alpha scores mean that it is appropriate, given also that other pre-requisites such as normal distribution were met, to analyse these data using conventional statistical techniques. The issue of rater reliability is discussed at more length in Section IV below, in which details of a post hoc inter-rater reliability study are given.

2. Standardisation of Test Procedures

One of the concerns that needed to be addressed in the design and application of the oral assessment instruments was the need for standardisation, particularly since it was planned to use a large number of assessors. Both local teachers and native English-speaking teachers from the schools within the sample group were invited to attend

training workshops to train as oral assessors. Those wishing to participate in the scheme were only allowed to conduct oral assessments on condition that they attended such a standardisation workshop in which test procedures, interview techniques and the application of the assessment criteria were carefully explained and discussed. Due to teacher availability, it was necessary to conduct a number of these workshops during which the assessors' test file packs were distributed, along with the audiotapes for recording. The assessment process and procedure were fully explained and discussed and carefully selected samples of a wide range of students' oral production (video taped during test development and trialling by the MENETS team) were shown to those attending the workshops. Participants were invited to apply the assessment criteria to students in videoed interviews and through discussion and the viewing of further videos, a consensus on students' level and the application of the criteria were arrived at. Assessors were also made aware of the fact that the audiotaping of the interviews was required for moderation and standardisation purposes and also for the second stage analysis.

The standardisation process was conducted prior to the first administration of the oral instruments and repeated before the final administration of the instruments for both first and second cohort groups. In addition, a CDROM was produced setting out the assessment procedures and providing examples of student performances. These performances provided a permanent reminder of the interpretations of the assessment criteria at each of the various levels.

VI. POST HOC INTER RATER RELIABILITY STUDY

A post hoc inter rater reliability study was conducted in order to analyse the extent to which trained raters concurred on the scores awarded to students who took part in this study. The three raters involved were all well trained and familiar with the interview procedure as well as the criteria used for scoring. On two separate occasions, audio tapes containing random samples of interviews were given to the raters who were asked to listen to the interviews once only, with no stopping or replaying. On the first occasion, a small random sample of ten students were selected and raters listened to the tape and conducted their grading independently. Correlation coefficients of this first round are shown on table 9 below.

Table 9. Correlation matrix of raters' oral assessment scores in inter-rater reliability study (round one)

	Rater 3	**Rater 1**
Rater 1	.91**	
Rater 2	.65*	.70*

** p = < 0.001 * p = < 0.05

All correlations were significant at the p = < 0.001 level but there was some concerns on the part of the raters regarding the poor sound quality of the tape being used (thus raising questions of reliability). Nevertheless, it was felt that the correlations were sufficiently high and a second round was conducted using a larger sample of sixteen students in which the same procedure that was employed in the first round was again

followed. Results of this second round of inter rater correlations are shown in table 10 below.

Table 10. Correlation matrix of raters' oral assessment scores in inter-rater reliability study (round two)

	Rater 3	**Rater 1**
Rater 1	.71**	
Rater 2	.84***	.85***

*** p = < 0.001 ** p = < 0.01

In this second round, all correlations were again significant at the p = < 0.001 level. The raters reported good sound quality of the audio tape and in comparison to the round one results, the correlation coefficients had a smaller range (.71 to .85) and were on average higher. In light of the results of this post hoc inter rater reliability study, there is good evidence to support the reliability of the assessment scores used in this study, although as discussed in chapter 8, such a study should ideally be conducted concurrently whilst the rater training is in progress during the course of the project.

A comparison of means between the three assessors was also conducted to ensure that in addition to the rank ordering of the assessors being similar, the means themselves were also similar (i.e. none of the assessors were too lenient/harsh). A series of paired-samples t-tests was thus conducted, with the results shown in the table 11 below.

Table 11. Inter rater reliability study. Comparison of assessors' mean scores

Assessor	mean	SD	Assessors	t-test t	sig
No 1	18.5	7.07	1 & 2	2.07	p = 0.07
No 2	15.0	3.26	2 & 3	-0.38	p = 0.71
No 3	15.3	2.00	1 & 3	1.90	p = 0.09

As we can see from table 11 above, although the mean score of assessor number one was higher than both of the other two, the t statistics and significance figures for all pairs indicate that the means of the three assessors were similar at the $p = < 0.05$ level. It should be noted however that despite the fact that mean score of assessor no. 1 (18.5) and assessor no. 2 (15.0) are not significantly different at the 0.05 level, since the maximum possible score is 30 this represents a difference of some 11% which is quite large. The literature suggests that this percentage of difference could be reduced with additional rater training and moderation processes.

Having discussed and detailed the methodology, instrument design and procedures employed in this study, the thesis now moves on to give details of the method and results of the data analysis which can be found in the following chapters.

Footnotes

[1] These assessment procedures were modeled along the lines of those commonly used by renowned language testing organisations, such as UCLES (University of Cambridge Local Examination Syndicate).

[2] This is a commonly recognized 'communication gap' exercise familiar to teachers and learners of the communicative classroom. (Ur, 1981).

[3] This second stage analysis was conducted by key members of the MENETS project team and is the subject of on-going research centred on error analysis and discourse analysis. One commonly accepted technique, for example, is to consider the Mean Length of Utterance (MLU) (MENETS, 2000).

[4] The assessment procedure that was developed involved training the assessors to make language choices in their questioning that moved from open question types (e.g. 'tell me about this') down to more closed questions if students were unable to respond (e.g. choice questions/yes-no questions). Wang and Gray (2001).

[5] MENETS (1999) point out that the reluctance of non-NET schools to become involved was understandable if regrettable. These schools may for example have made a conscious decision not to take advantage of the NET funding for some specific reason and this might in any case have affected the usefulness of having the school in the sample.

CHAPTER 4

PRELIMINARY DATA ANALYSIS

I. DATA ANALYSIS OF PRE TEST AND POST TEST SCORES

Since the primary focus of this study is to determine the impact of NET teachers on the pupils' increase in English language proficiency, the data from the pre test and post test was subjected to analysis consisting of a number of different stages, namely:

i. Descriptive analysis

ii. Fixed point in time analysis of variance (ANOVA)

iii. Rasch Scale Modelling (RSM)

iv. Ordinary least squares (OLS) multiple regression

v. Multi Level Modelling (MLM) of the Rasch calibrated scores

This chapter sets out to provide the rationale for the analysis, describe the various different stages involved, reveal the results of these analyses and finally to arrive at some tentative conclusions as regards the implications and inferences that can reasonably be made from these findings.

1. Descriptive analysis

> "Description lays the basis for analysis, but analysis also lays the basis for further description. Through analysis, we can obtain a fresh view of our data ... The core of quatitative analysis lies in these related processes of describing phenomena, classifying it, and seeing how our concepts interconnect" (Dey, 1993:30).

Following the two administrations of the oral proficiency test (i.e. pre and post tests), the data was first of all 'cleaned' of errors and then a descriptive analysis was conducted on the different variables on the data base. This was intended to help the researcher determine amongst other things the validity of this type of statistical analysis, since standard statistical procedures are based on a number of assumptions underlying the data. There are assumptions for example that the distribution is more or less normal, and if this is the case then the range and standard deviations will be within acceptable limits and the distribution will not be positively or negatively skewed. 'Normal' in this case is used to describe a unimodal, bell shaped curve which is symmetric about its mean. In a standard normal distribution, the population mean (μ) equals zero and a standard deviation (σ) of one. In such a distribution, 68.26% of the observations fall between $\pm$ one standard deviation and 95.44% of the observations fall between $\pm$ two standard deviations (Hopkins et al, 1996). Other factors such as the measures of stability (or sampling error) as reflected in the standard error figures also need to be considered since anomalies in any of these areas can threaten the validity of the subsequent statistical analysis.

2. List of Variables

The final list of variables on the data base consisted of the following:

i) *School Level (SCH_LV):* At the time of data analysis, secondary schools in Hong Kong were divided into five bands[1], with band 1 being the "highest" and band 5

being the "lowest". Pupils leaving primary schools were allocated a secondary school according to a complex testing system, the Secondary School Placement Allocation System (SSPAS), whereby the top 20% of pupils were allocated to band 1, the next 20% to band 2, and so on. For the purposes of this study, the banding was simplified and schools were designated either 'high', 'medium' or 'low' according to information supplied to the MENETS project by the Hong Kong Education Department.

ii) *Pupils Form (FORM1):* In this study, three groups of pupils were being investigated, Form 1, Form 3 and Form 4. It was necessary to differentiate between the three different forms since the impact of a NET teacher might for example be different at different ages/forms.

iii) *Unique identification (ID):* For reasons that are self-explanatory, it was necessary to allocate a unique identification for each pupil so that further statistical analysis could be conducted.

iv) *Student Level in First Year (STU_LV1):* According to class performance, formal and informal test results and home work assignments, teachers were asked to categorise pupils as high, medium or low ability. The purpose was to investigate any possible NET effect on pupils of different ability for the first year of the study (1998-1999).

v) *Student Level in Second Year (STU_LV2):* This variable is the same as the previous one except that the teachers' categorisations are for the second year of the study (1999-2000).

vi) *Pre Test Score (ORAL1):* Total raw score of the pre test (Time 1) measurement of pupils' oral English language proficiency. This total raw score in turn consisted of five sub scores as follows: *Comprehension (COMP1); Fluency (FLU1); Vocabulary Resource (VOC1); Grammatical Accuracy (GRAM1); Pronunciation (PRON1)* (see Chapter three, III.2 for full details).

vii) *Post Test Score (ORAL2):* Total raw score of the post test (time 2) measurement of pupils' oral English language proficiency. This again consisted of the same five sub scores *Comprehension (COMP2); Fluency (FLU2); Vocabulary Resource (VOC2); Grammatical Accuracy (GRAM2); Pronunciation (PRON2).*

viii) *Rasch Calibrated Scores of Pre Test (RSM1):* Total raw scores of the pre test were calibrated in Rasch Scale Modelling to produce adjusted Rasch calibrated scores. These Rasch scores were then used in the Ordinary Least Squares and Multi Level Modelling that was subsequently conducted.

ix) *Rasch Calibrated Scores of Post Test (RSM2):* Calibrated Rasch scores of the raw post test scores.

x) *School District (DIS):* In order to help ensure randomised stratified data, it was necessary to differentiate between schools in the three different geographical areas i.e. New territories (NT); Kowloon (KLN); Hong Kong island (HK).

xi) *Pupil's Teacher in First Year (T98_99):* Since the primary aim of this study is to investigate any possible "NET effect", it was necessary to distinguish between those pupils taught a) only by NETs, b) only by local teachers, and c) by a combination of NETs and local teachers.

xii) *Pupil's Teacher in Second Year (T99_00):* Since this was a two year study, it was also necessary to identify the pupils' English teacher(s) in the second year.

Although the categories xi) and xii) above appear to produce only three basic categories, 'NET', 'Local' and 'Both', the situation is more complex. Over the two-year period of study were are in fact nine different possible variations of teaching mode as shown below:

Different possible combinations of teaching mode.

Groups	In the 1st (1998-1999) school year, the group was taught English by:	In the 2nd (1999-2000) school year, the group was taught English by:
1	a NET	a NET
2	a NET	a LOCAL English teacher
3	a NET	BOTH a NET and a LOCAL English teacher
4	a LOCAL English teacher	a LOCAL English teacher
5	a LOCAL English teacher	BOTH a NET and a LOCAL English teacher
6	a LOCAL English teacher	a NET
7	BOTH a NET and a LOCAL English teacher	BOTH a NET and a LOCAL English teacher
8	BOTH a NET and a LOCAL English teacher	a LOCAL English teacher
9	BOTH a NET and a LOCAL English teacher	a NET

In addition, there was an extra complexity with the category 'BOTH' (i.e. students taught by a combination of NET teacher *and* local teacher). The difficulty arose because this category included a wide range of combinations in which the NET-local teacher ratio (percentage) varies. In addition, where classes were taught by 'both', the NET teachers' responsibilities often differed. For example, some NETs taught speaking skills only, some taught listening *and* speaking and some split the classes 50-50 with the local teacher with the half groups changing over at half term. Such was the complexity

of these arrangements that it was not possible to strictly control the variable 'both'. An initial attempt was made to quantify the number of hours that students were taught be the NET over the whole academic year but the collection of this data from schools was not readily forthcoming and rather than risk loosing further, unacceptable amounts of data, this idea was abandoned. In the end, group 1 (above) became the 'NET' category, group 4 (above) became the 'Local' category and the remaining seven groups, which as we have seen within itself contains considerable variation, became the 'both' category.

Having considered and discussed the different variables involved in this study, the next section considers the initial descriptive analysis conducted in this study.

II. ANALYSIS OF RAW ORAL ASSESSMENT SCORES

1. Whole Sample

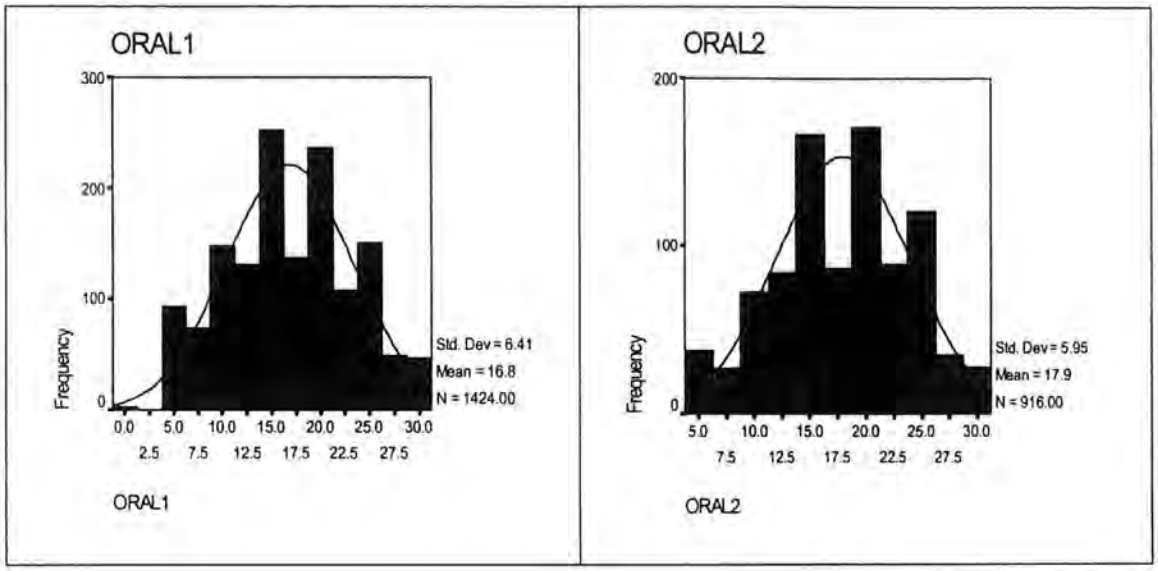
An analysis of the raw scores of the two administrations of the oral assessment (pre test/Oral1 and post test/Oral2) was first of all conducted and the resulting descriptive statistics suggest that the two distributions were more or less normal. A summary of the statistics is given in table 12 below:

Table 12. Summary statistics of raw oral assessment scores

	ORAL1	ORAL2
n - valid	1424	916
missing	0	508
Mean	16.82	17.95
S. E. of Mean	0.17	0.19
Median	17.00	18.00
Mode	15.00	20.00
S D	6.41	5.94
Skewness	-0.001	-0.144
Kurtosis	-0.73	-0.61
Range	30.00	25.00
Minimum	0.00	5.00
Maximum	30.00	30.00

The above statistics indicates that the data may be analysed using standard traditional statistical procedures since the basic assumptions regarding this type of analyses have largely been met. In both pre and post tests, the Skewness of -0.001 and -0.144 are well within accepted limits of $\leq \pm 1$ although the Kurtosis statistics of -0.727 and -0.614 are somewhat larger, they are again well within accepted limits of $\leq \pm 1$ (e.g. Dey, 1993; Morgan et al, 2001). The standard deviations of 6.412 and 5.94 again suggest a dispersed and more or less normal distribution. Figure 1 below shows the histogram of the pre and post tests with normal distribution curves.

Figure 1. Histograms showing distribution of pre and post test raw scores



It is interesting to note even from this preliminary analysis that there was a tendency for examiners to mark holistically rather than use the descriptive criteria to discriminate more finely between pupils. In the pre test, for example the score of 5 was awarded to 63 pupils (4.4%), 10 to 54 pupils (3.8%), 15 to 89 pupils (6.3%), 20 to 83 pupils (5.8%), 25 to 48 pupils (3.4%) and 30 to 30 pupils (2.1%). Thus in nearly 30% of all scores, examiners tended to give the same mark for the five different and supposedly discrete areas of comprehension, vocabulary resource, fluency, grammatical accuracy and pronunciation. Post test scores were very similar in this respect. This apparent lack of discrimination between the five oral assessment criteria is an issue that we will return to later, since in subsequent analysis (e.g. Rasch scale modelling) the statistical ramifications have important consequences.

Inter-item correlations and Alpha scores were calculated for both pre and post tests with the results shown in the tables below:

Table 13. Correlation matrix of items in oral assessment pre test

	Comprehension	Fluency	Vocabulary	Grammatical accuracy
Fluency	.87			
Vocabulary	.86	.89		
Grammatical accuracy	.84	.88	.89	
Pronunciation	.82	.85	.86	.88

The standardised item alpha of $\alpha = 0.97$ was considered to be very high, and the figure could not have been increased by deleting any of the items. This again suggests that the five items were not discriminating between the five areas and that examiners were marking holistically. The post test inter-item correlations were very similar as can be seen in table 14 below:

Table 14. Correlation matrix of items in oral assessment post test

	Comprehension	Fluency	Vocabulary	Grammatical accuracy
Fluency	.86			
Vocabulary	.82	.87		
Grammatical accuracy	.81	.86	.88	
Pronunciation	.80	.84	.84	.85

The standardised item alpha of $\alpha = 0.97$ was again high and could not be increased by deleting any of the items. Although the internal consistency reliability was

high in both administrations, it is not suggested that this was in fact a true indication of the reliability of the tests which is more likely to be a measure of the inter rater reliability of the assessors (see Chapter 3, p. 68).

Two further statistical tests of normality of the pre test and post test raw scores were conducted, namely the Kolmogorov-Smirnov test of normality and the Shapiro-Wilk test of normality (e.g. Pallant, 2001) with the following results:

Table 15. Kolmogorov-Smirnov and Shapiro-Wilk statistics on normality

Pre Test	
Kolmogorov-Smirnov statistic:	0.051 (df, 1424), p = <0.001
Shapiro-Wilk statistic:	0.983 (df, 1424), p = <0.001
Post Test	
Kolmogorov-Smirnov statistic:	0.076 (df, 916), p = <0.001
Shapiro-Wilk statistic:	0.984 (df, 916), p = <0.001

Since the above statistics are significant at the $p = < 0.01$ level, this indicates some deviation from normality. Despite the significance of these two tests of normality, subsequent analyses of the distributions were robust as we can see in the following sections. It was felt that overall the distributions could be considered as more or less normal and the data could be analysed using statistical techniques that assume normality.

The descriptive analyses described above were carried out on the two sample groups *as a whole*, i.e. 1424 and 916 students for the pre and post tests respectively. In

addition, these two sample groups were further divided into the three age/year groups (form 1, form 3 and form 4) and the distributions were further analysed to determine whether the resulting sub-sets also conformed to the assumptions regarding normal distributions. The following section reports on the distribution of forms one, three and four respectively.

2. Analysis By Age/Form

Form One

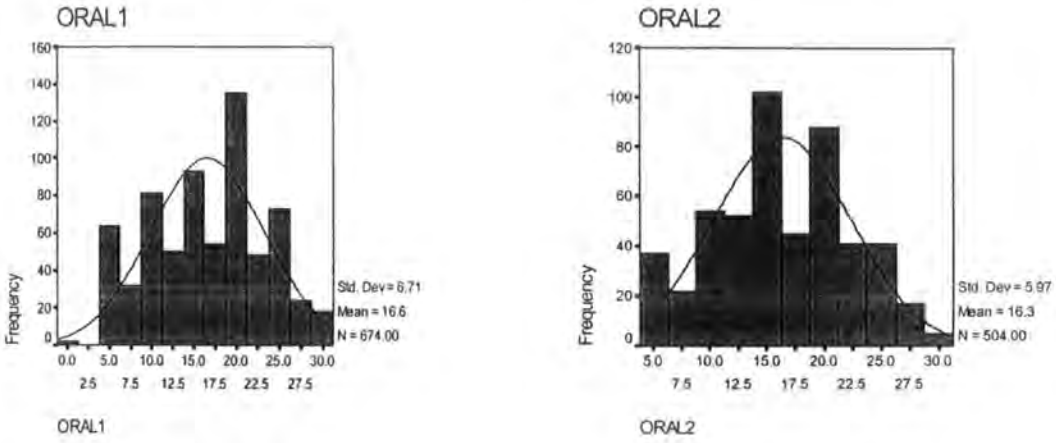
Following the descriptive analysis of the whole sample, the data was then analysed by age/form. The results of the form one analysis are summarised in table 16 below.

Table 16. Summary statistics: form one pre and post tests

	Pre Test	Post Test
N - Valid	674	504
Missing	16	186
Mean	16.56	16.33
S. E. Mean	0.26	0.26
Median	17.00	16.00
Mode	5.00	20.00
SD	6.71	5.97
Skewness	-0.098	-0.011
Kurtosis	-0.858	-0.666
Range	30.00	25.00
Minimum	.00	5.00
Maximum	30.00	30.00

As with the whole samples, the distribution is more or less normal and the statistics seem to be within acceptable limits as can be seen from Figure 2 below.

Figure 2. Histograms of pre and post test scores of form one students



Form Three

The form three students were then analysed and the results of the descriptive analysis are shown in table 17 below.

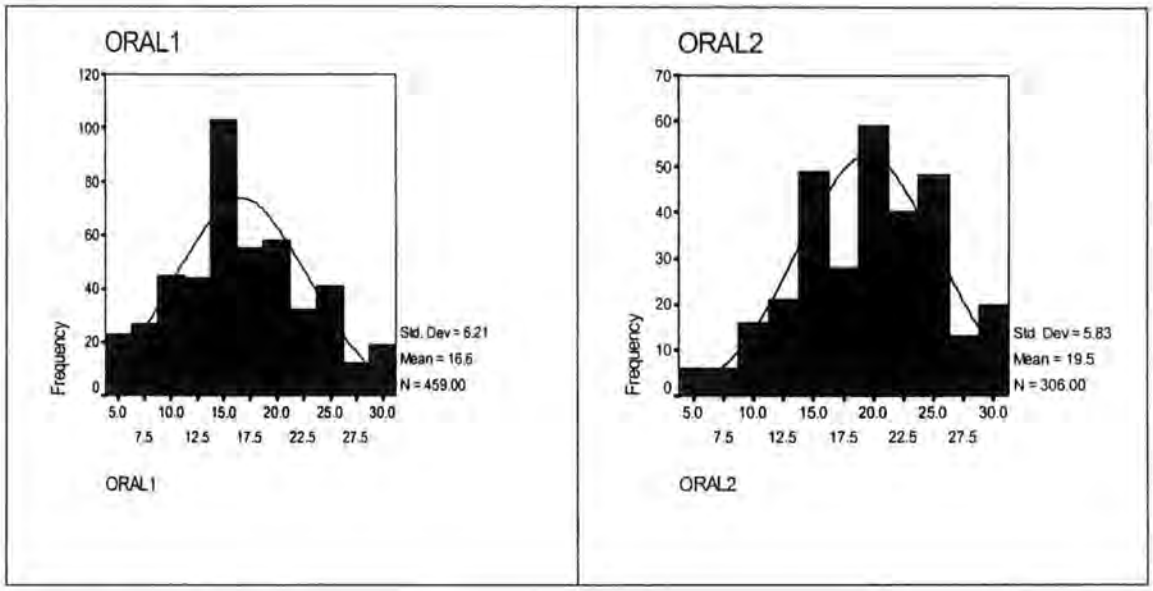
Table 17. Summary statistics: form three pre and post tests

	ORAL1	ORAL2
N - Valid	459	306
Missing	17	170
Mean	16.64	19.47
Median	16.00	20.00
S D	6.20	5.83
Skewness	0.201	-0.270
Kurtosis	-0.557	-0.434
Range	25.00	25.00
Minimum	5.00	5.00
Maximum	30.00	30.00

The above statistics again show that the scores of this sub group are more or less normally distributed in both the pre and post tests with skewness (pre test 0.201; post test -0.270) and kurtosis (pre test -0.557; post test -0.434) within the acceptable limits of ±1.

The histograms in figure 3 below show the distributions of form three students in the pre and post tests and again suggest that the respective distributions are more or less normal and can further analysed using standard statistical techniques that are based on the assumption of a normal distribution.

Figure 3. Histograms of pre and post test scores of form three students



Form Four

In the case of the form four students, the kurtosis of -1.07 in the post test is a violation of one of the assumptions underlying a normal distribution and thus we should advise caution in the interpretation of further analysis of this set of data. However, this figure is not far in excess of ±1 and since all the other data sets are within the generally accepted limits, it was felt that further analysis could proceed. Table 18 below shows the summary statistics of form 4 pre and post test oral assessment scores.

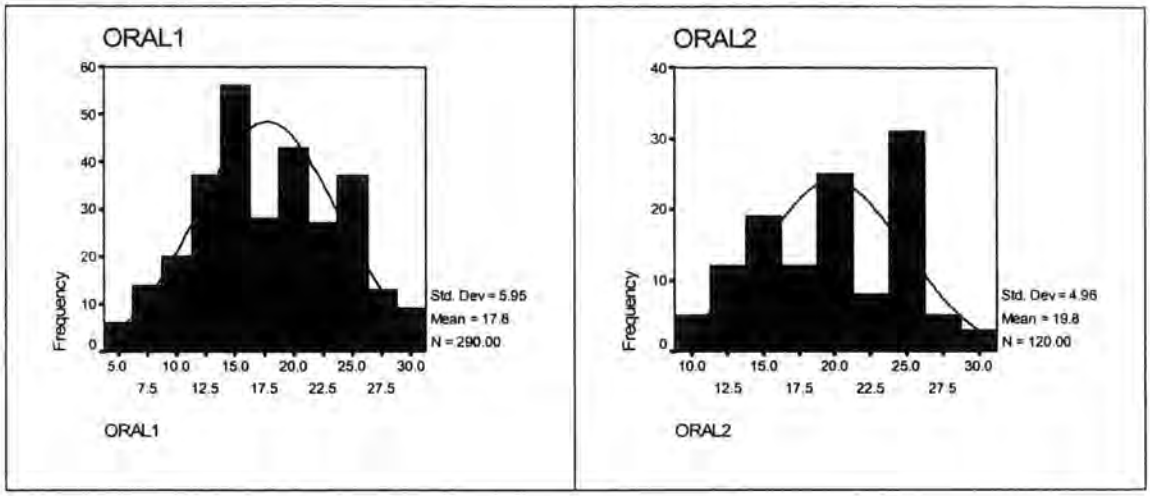
Whilst the skewness of both the pre test and the post test are within acceptable limits (0.04 and -0.13 respectively), the kurtosis statistics suggest that the shape of the distribution curve is not normal (pre test -0.76, post test -1.07) which is illustrated in the histograms in Figure four below.

Table 18. Summary statistics: form four pre and post tests

	ORAL1	ORAL2
n - valid	290	122
missing	2	170
mean	17.76	19.77
median	18.00	20.00
S D	5.95	4.96
skewness	0.04	-0.13
kurtosis	-.76	-1.07
range	25.00	18.00
minimum	5.00	11.00
maximum	30.00	29.00

The histograms in Figure four below of pre and post test scores of form four students show a somewhat 'flat' distribution, particularly in the latter. The slight negative skew is also apparent in the histogram, as is the fact that some 8.6% of the observations are below minus two standard deviations from the mean (in a normal distribution this would only be 2.28%), and 24.2% of the observations are below minus one standard deviations from the mean (in a normal distribution this would only be 15.87%). Since there is some evidence that the form four post test scores vary somewhat from a normal distribution, we should be cautious in the subsequent analysis of these data.

Figure 4. Histograms of pre and post test scores of form four students



This chapter will now present and discuss the preliminary, descriptive analysis conducted on the other variables in the data set.

III. DESCRIPTIVE ANALYSIS OF OTHER VARIABLES

1. School Level

In this study, a total of 46 schools were involved and all were classified as either 'high', 'medium' or 'low' banding according to information supplied to the MENETS project by the Hong Kong Education Department (ED). The distribution of scores according to school level was as follows:

Table 19. Oral assessments: school level

	School level (n)	student assessments (n)
Low	11 (23.91%)	317 (22.3%)
Medium	20 (43.47%)	631 (44.3%)
High	15 (32.61%)	476 (33.4%)
TOTAL	46	1424

2. School District

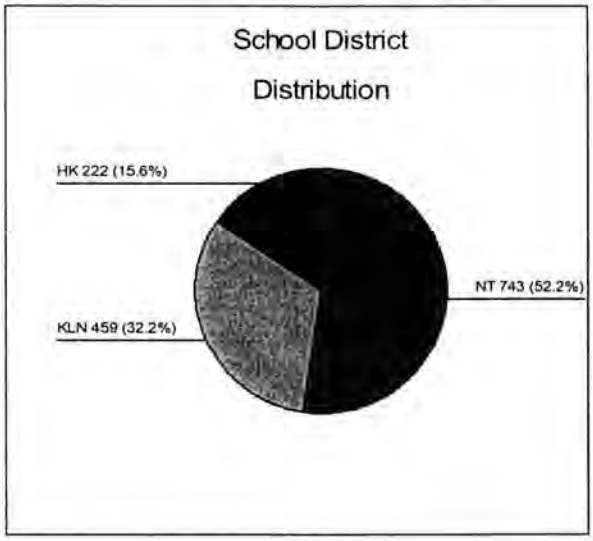
The research methodology of the MENETS project was designed to ensure a randomised, stratified sample of oral assessments. It was thus necessary to ensure that not only was the sample school level representative of the population, but also that the schools in the sample came from geographical locations that were representative of Hong Kong as a whole i.e. that the sample represented all three areas, namely New Territories, Kowloon and Hong Kong island. A summary of school areas is shown in table 20 below:

Table 20. Oral assessments: school district

	total schools (n)	student assessments (n)
NT	23 (50.0%)	743 (52.2%)
KLN	16 (34.78%)	459 (32.2%)
HK	7(15.21%)	222 (15.6%)
TOTAL	46	1424

The pie chart in figure 5 below shows the distribution of oral assessment scores by school district. As we can see, over fifty percent of the observations took place in the New Territories, with Kowloon being the second largest group (thirty five percent) and Hong Kong the smallest of the school groups (fifteen percent). This is in line with the randomised, stratified sampling designed to reflect the constituency of schools throughout the Hong Kong SAR.

Figure 5. Pie chart showing distribution of oral scores by school district



This section has covered the descriptive analysis of the data which includes Time one and Time 2 administrations of the oral English proficiency assessment. With some exceptions, this descriptive analysis suggests that both administrations have more or less normal distributions, indicating that standard statistical techniques would be appropriate and that such an analysis would enable us to make inferences as to any possible 'NET effect' on students' proficiency gain between time one and time two. In addition, the

statistical techniques that will be described and discussed in subsequent chapters also throw light on the impact of other key variables in this proficiency gain.

The following section continues the data analysis, focusing on the comparison of different means through the use of analysis of variance (ANOVA) and paired sample t-tests. Although these comparisons of means are not the main foci of this investigation, they are nevertheless important in helping us to arrive at an understanding of the factors affecting oral English language proficiency gain.

IV. COMPARISONS OF MEANS

This section of the thesis presents the preliminary analysis of the oral English assessment scores by analysing the respective means of different groups of students, such as those being taught in English medium as opposed to Chinese medium schools. As mentioned in Chapter two, previous studies such as the interim report on the EELTS pilot scheme (British Council, 1988) focused almost exclusively on analysing the means of different sub-groups and making inferences from any observed significant differences. The shortcomings of this type of analysis have already been alluded to but are based on the observation that such an analysis considers time one and time two to be two discrete points in time rather building a regression model which considers the two means simultaneously. Nevertheless, a similar analysis of means was conducted on data arising from this research, not least because variables showing significant differences between

groups are likely to be important indicators that should be explored in greater depth in later modelling. Another reservation on the analysis conducted to date in similar research lies in the fact that variables, although technically independent, can impact on each other. In other words there is often some interaction between two or more variables involved in a study and this interaction can not be explored through restricting the analysis to conducting a series of t-tests in which the mean scores are considered to be independent. We will look firstly at the overall pre and post test scores.

1. Pre test and Post Test Comparisons: Time 1 to Time 2 Gain

A paired sample t-test was conducted to evaluate whether there was a significant gain in the oral assessment scores of the whole sample between the pre and the post test scores. Given a reliable assessment instrument, such a gain in scores is what one might reasonably predict. The t-test revealed that there was a statistically significant increase in the assessment scores between time 1 ($M = 16.78$, $SD = 6.30$) and time 2 ($M = 17.95$, $SD = 5.95$), $t(915) = -7.04$, $p < 0.01$**. The eta squared statistic (0.05) indicates a small to moderate effect size (Cohen, 1988).

Similar paired samples t-tests were carried out on the three different age/year groups (forms 1, 3 and 4), with the following results:

Form 1: There was no statistically significant increase in the assessment scores of form one students between time 1 ($M = 16.40$, $SD = 6.49$) and time 2 ($M = 16.55$, $SD =$

5.88), t (915) = -0.68, p = 0.493. There was no measurable effect size for this group (eta squared statistic was close to zero at .0005).

Form 3: A similar paired samples t-test was carried out on the form three students where there was a statistically significant increase in the assessment scores between time 1 (M = 17.13, SD = 6.28) and time 2 (M = 19.44, SD = 5.86), t (305) = -7.69, $p < 0.01$**. For form three students, the eta squared statistic of 0.16 indicates a large effect size.

Form 4: Finally, a similar paired samples t-test was carried out on the form four students where again there was a statistically significant increase in the assessment scores between time 1 (M = 17.40, SD = 5.49) and time 2 (M = 19.83, SD = 4.96), t (119) = -6.41, $p < 0.01$**. The eta squared statistic (0.26) again indicates a large effect size.

The above results tend to support the commonly held view that in Hong Kong when students leave primary school and enter secondary school their performance over a wide range of subjects falls. This phenomenon is especially true of those students entering schools where English is the medium of instruction (EMI). The reason commonly cited is that where previously students had only studied English as a subject, in an EMI school all their other subjects are also taught through English. This presents students with a number of cognitive and linguistic challenges, such as the high intensity of new subject-specific lexis, decoding the teacher's oral input, etc. It is generally felt that this 'dip' in the learning curve is compensated for by the time students reach form two. Thus the lack of gain between Time 1 and Time 2 assessment scores in form one

students detailed above is consistent with generally held views on students' transition from primary to secondary school. Also consistent with this are the findings of the t-tests for form three and form four students where significant gain was observed.

2. Pre Test Analysis

A one-way between-groups analysis of variance (ANOVA) was conducted on the pre-test oral assessment scores (Oral 1) in order to compare the school levels. Here again there were found to be statistically significant differences at the $p<.01$ level between all three different groups of high, medium and low level schools [$F(2, 1421) = 261.01$, $p=<0.01$**]. Post hoc comparisons were again calculated using the Tukey HSD test, showing that the mean scores of all three schools levels were significantly different from each other, as follows: high ($M = 20.86$, $SD = 5.23$); medium ($M = 16.27$, $SD = 5.63$); low ($M = 11.87$, $SD = 5.77$). The eta squared statistic was 0.27, indicating a large effect size.

The above analyses which were conducted on the whole student sample in which the results of form one students were combined with those of form three and form four students. As with the paired-samples t-test discussed above, the data analysis then considered the three populations independently and the results are outlined in the section below.

Form One Students

A one-way between-groups analysis of variance (ANOVA) was also conducted separately on the form one students and statistically significant differences at the p<.01 level were observed between all three different groups of high, medium and low level schools [F(2, 672), = 168.96, p=<0.01**]. As with the whole sample analyses, post hoc comparisons were calculated using the Tukey HSD test. This showed that the mean scores of all three schools levels were significantly different from each other, as follows: high (M = 20.61, SD = 5.14); medium (M = 16.44, SD = 5.81); low (M = 10.69, SD = 5.46). The effect size, calculated using eta squared, was 0.33.

Form Three Students

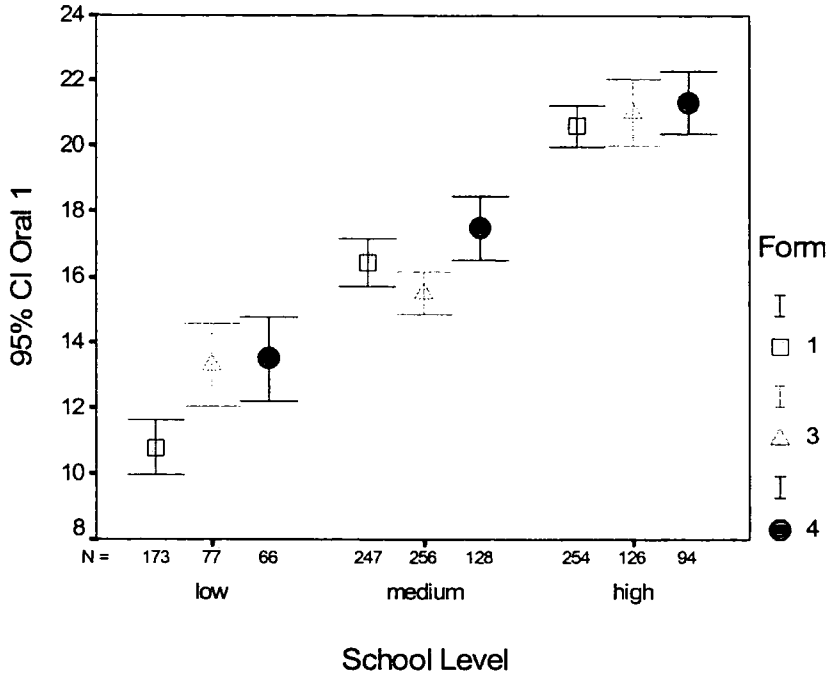
Similarly, an ANOVA conducted separately on the Form three students showed statistically significant differences at the p<.01 level between all three different groups of schools [F(2, 456), = 58.98, p=<0.01**]. Post hoc comparisons using the Tukey HSD test showed a significant difference between the mean scores of all three groups, as follows: high (M = 21.03, SD = 5.814); medium (M = 15.49, SD = 5.39); low (M = 13.30, SD = 5.59). The effect size, calculated using eta squared, was again high at 0.21.

Form Four Students

Similar results were noted from a one-way analysis of variance conducted separately on the form four students. Again, statistically significant differences at the $p<.01$ level between all three different groups of schools were noted [$F(2, 278)$, = 47.52, $p=<0.01^{**}$]. A Tukey HSD post hoc comparison test showed significant differences between the mean scores of all three groups: high (M = 21.31, SD = 4.63); medium (M = 17.48, SD = 5.55); low (M = 13.29, SD = 5.17). The effect size, calculated using eta squared, was 0.25 was considered to be large.

Figure 6 below gives an error bar summary graph of pre test oral assessment scores for forms one, three and four and in high, medium and low level schools. Whilst there is a clear difference *between* scores in the three levels of school, the difference between the three forms is less noticeable. In low level schools the scores of Form one students, are clearly different from those of Form three and four students. In medium and high level schools, the differences are less pronounced.

Figure 6. Error bars showing confidence intervals of pre test scores for forms one, three and four at different school levels



Even though by this early stage of the research, the students' exposure to NET teachers was minimal, it is clear that the school itself and in particular the level of school is strongly related to the oral assessment scores and this school variable is clearly one that needs to be explored in greater depth in the data analysis. This is true not only in forms three and four as one might predict, but also in form one where students have recently arrived from primary school.

3. Post Test Analysis

Whole sample

A one-way between-groups analysis of variance (ANOVA) was conducted to explore the possible impact of school level/banding on the post test oral assessment

scores (Oral 2). Students test scores were analysed according to whether they were studying in a high, medium or low level school. As we might predict, given a reliable test instrument, there were statistically significant differences at the p<.01 level in Oral 2 scores for the three different groups [F(2, 913) = 135.24, p=<0.01**]. Post hoc comparisons using the Tukey HSD test indicated that the mean scores for all three schools levels were significantly different from each other, as follows: high (M = 20.93, SD = 5.16); medium (M = 17.77, SD = 5.21); low (M = 13.58, SD = 5.37). The effect size, calculated using eta squared, was 0.23, indicating a large effect size.

Form One students

As with the whole sample, statistically significant differences in an ANOVA were observed at the p<.01 level in Oral 2 scores for form one students [F(2, 486) = 98.13, p=<0.01**]. A Tukey HSD post hoc comparison indicated that the mean scores for all three schools levels were significantly different from each other: high level (M = 20.00, SD = 5.14); medium level (M = 16.29, SD = 4.72); low level (M = 11.85, SD = 5.09). The effect size, again calculated using eta squared was 0.29.

Form Three students

Statistically significant differences were also observed from an ANOVA at the p<.01 level in Oral 2 scores for form three students [F(2, 304) = 53.40, p = <0.01**]. The mean scores for all three schools levels were significantly different from each other

as revealed by a Tukey HSD post hoc comparison. The results were as follows: high level (M = 22.17, SD = 5.08); medium level (M = 19.65, SD = 5.28); low level (M = 14.00, SD = 4.50). The eta squared effect size of 0.26 is considered to be large.

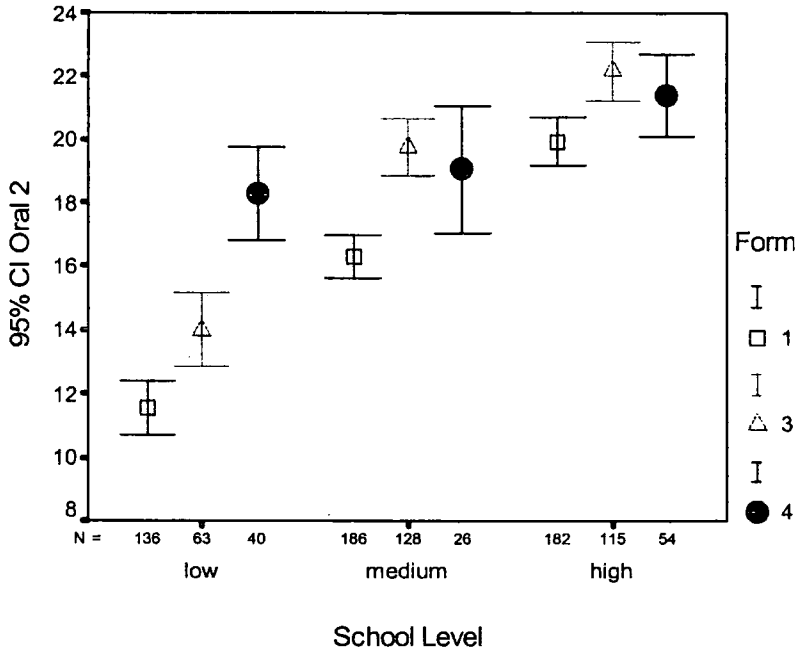
Form Four Students

Statistically significant differences were also observed as a result of an ANOVA conducted on the Oral 2 scores of form four students at the $p<0.01$ level [$F(2, 117) = 5.43$, $p = <0.01^{**}$]. A Tukey HSD post hoc comparison test indicated that the mean scores of high level schools (M = 21.39, SD = 4.82) were significantly higher than those of low level schools (M = 18.22, SD = 4.61). However, the mean scores of students in medium level schools (M = 19.07, SD = 4.96) were not significantly different from those in either high or low level schools. Eta square was used to calculate the effect size which was moderate at 0.08.

Figure 7 below gives an error bar summary graph of post test oral assessment scores for forms one, three and four and in high, medium and low level schools. The scores of Form one and Form three students are clearly different according to the level of school. However, in the case of Form four students' scores, those in low level schools do not show a noticeable difference from those in medium level schools. The Form four students' scores in medium and high level schools are not as high as we might have predicted. However, the cell numbers of form four students are much smaller than those

of form one and form three students. In addition, the variation of scores in form four students in medium level schools is greater than that on the other sub-groups.

Figure 7. Error bars showing confidence intervals of post test scores for forms one, three and four at different school levels



4. Medium of instruction

A one-way analysis of variance was conducted to explore the relationship between students' oral assessment scores and their medium of instruction in other subjects. It is argued that those students in EMI schools will be exposed to more English and have more opportunity to communicate through the L2 than their counterparts in CMI schools and as a consequence they are likely to show more gain in oral English language proficiency.

i) Pre Test

An ANOVA analysis showed that there was a statistically significant difference between the oral 1 scores of those students taught in EMI schools and those taught in CMI schools at the $p<0.01$ level [$F(1, 1422) = 318.21$, $p = <0.01^{**}$]. The mean score of students taught in EMI schools was $M = 20.50$, $SD = 5.27$ while the mean score of students taught in CMI schools was $M = 14.78$, $SD = 6.07$. The effect size (eta squared) was 0.18, indicating that the medium of instruction has a large effect on pre test outcomes when the whole sample is analysed together.

A similar analysis was conducted on the three different age/year groups independently, with the following results:

Form 1: $F(1, 673) = 202.22$, $p<0.01^{**}$

EMI: $M = 20.89$, $SD = 4.97$; CMI: $M = 14.16$, $SD = 6.32$

Effect size (eta squared) = 0.23

Form 3: $F(1, 457) = 80.56$, $p<0.01^{**}$

EMI: $M = 20.23$, $SD = 6.08$; CMI: $M = 15.04$, $SD = 5.56$

Effect size (eta squared) = 0.15

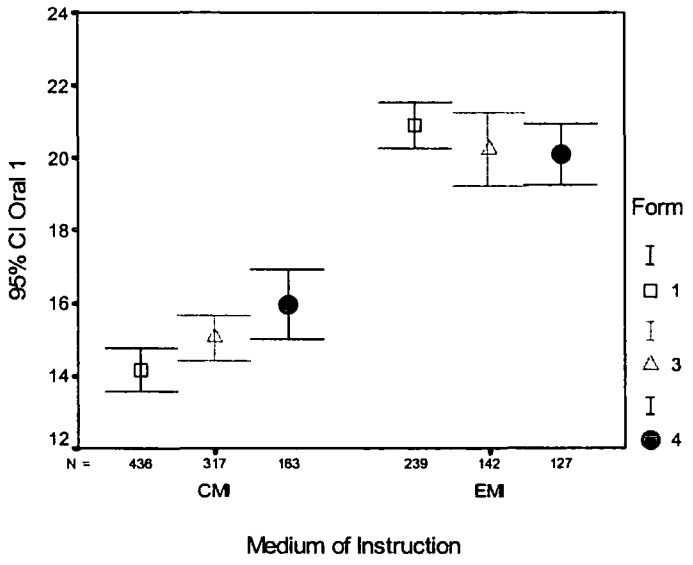
Form 4: $F(1, 288) = 38.76$, $p<0.01^{**}$

EMI: $M = 20.08$, $SD = 4.80$; CMI: $M = 15.96$, $SD = 6.15$

Effect size (eta squared) = 0.12

Figure 8 below illustrates the analysis conducted by form on medium of instruction and pre test oral assessment scores and we can clearly see the difference between the results in EMI and CMI schools. Interestingly, Forms one, three and four are clearly separated as we would predict in the CMI schools, but in the EMI schools all three forms have on average similar mean scores. This might be due to a 'levelling off' in EMI schools where all students' exposure to the English language is much greater, whereas in the CMI schools the only exposure is during English (as a subject) lessons.

Figure 8. Error bars showing confidence intervals of pre test scores for forms one, three and four in EMI and CMI schools



This preliminary analysis of the means of students taught in EMI schools as compared to their counterparts taught in CMI schools shows that the medium of instruction is an important variable, as one might predict, and will need to be explored further in the modelling stage of this data analysis.

ii) Post test

A similar ANOVA analysis was conducted on the oral 2 scores, with the following results:

Whole sample: $F(1, 914) = 225.23$, $p<0.01$**

EMI: $M = 21.38$, $SD = 4.99$; CMI: $M = 15.91$, $SD = 5.52$

Effect size (eta squared) = 0.20

Form 1: $F(1, 487) = 155.45$, $p<0.01$**

EMI: $M = 20.58$, $SD = 4.93$; CMI: $M = 14.48$, $SD = 5.22$

Effect size (eta squared) = 0.24

Form 3: $F(1, 305) = 64.77$, $p<0.01$**

EMI: $M = 22.47$, $SD = 4.97$; CMI: $M = 17.46$, $SD = 5.55$

Effect size (eta squared) = 0.17

Form 4: $F(1, 118) = 10.41$, $p = 0.002$**

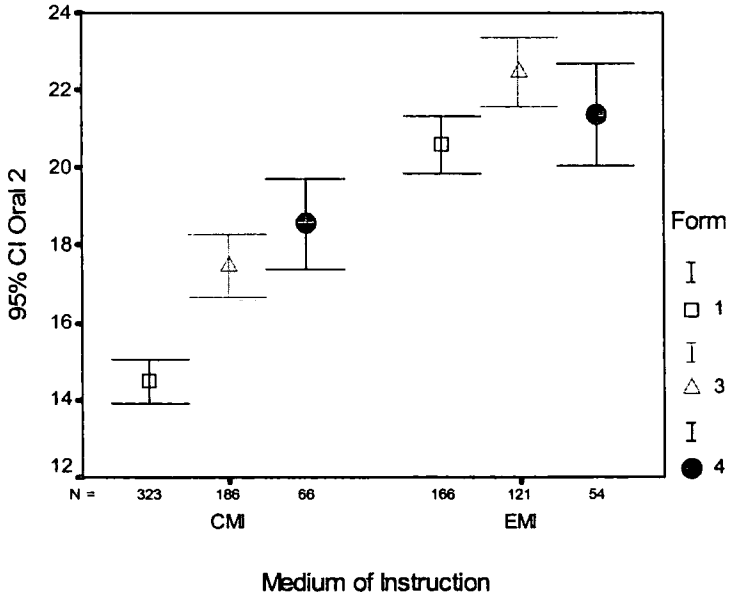
EMI: $M = 21.39$, $SD = 4.83$; CMI: $M = 18.56$, $SD = 4.74$

Effect size (eta squared) = 0.08

Figure 9 below illustrates the above analysis conducted by separate forms in different medium of instruction schools and post test scores and we can clearly see again

a distinct difference between the mean scores of the EMI students as opposed to the CMI students, although the gap seems to have narrowed compared to the pre test scores (see Figure 8 above). Once again, in the EMI schools there seems to be little difference between the three groups (once errors are taken into account) whereas in the CMI schools, there is a distinct difference between the mean scores of Form one and Form four students. Interestingly, there is again a difference in variance between the sub-groups, with the spread of the Form four EMI students being greater than that of the others.

Figure 9. Error bars showing confidence intervals of post test scores for forms one, three and four in EMI and CMI schools



As had been predicted, the results of this analysis show that the medium of instruction is a significant variable in all three year/age groups, with effect sizes ranging from large to very large[2]. These results support the theory that the medium of instruction has a strong relationship with English language oral assessment scores and suggest that this variable is an important one to consider in further regression and multi-level

modelling. It is interesting to note that even in form 1, when students have newly arrived at their secondary schools, the medium of instruction appears to have a strong relationship with students' outcomes in language proficiency.

5. Teaching Mode Analysis

A further series of analyses of means (ANOVA) was conducted to determine whether there were any significant differences in the mean scores of students taught by NETs, local teachers and both (NET and local teachers).

i) Pre Test and Teaching Mode

Whole Sample

An ANOVA carried out on the pre test scores of the whole sample reveal that there was a significant difference in the scores of students taught by NETs and the scores of those taught either by local teachers or by those taught by 'both' although the effect size is very small. Table 21 below gives full details.

Table 21. Summary ANOVA table of pre test scores according to teacher mode (whole sample)

Teacher	mean score	S.D.
Local	16.13	7.04
Both	16.96	5.43
NET	17.55	6.35
$F(2,1421) = 6.37$, $p = <0.01$**		
Effect Size = 0.009		

Form One

In the case of Form one students, there was a significant difference between the scores of students taught by local teachers and those taught by 'both'. In this case the effect size is moderate to large at 0.12. Full statistics are given in table 22 below.

Table 22. Summary ANOVA table of pre test scores according to teacher mode (Form one)

Teacher	mean score	S.D.
Local	15.62	7.28
Both	17.45	5.48
NET	16.78	6.86
$F(2, 672) = 4.22$, $p = < .05$*		
Effect Size = 0.12		

Form Three

With Form three, the mean scores of those students taught by NETs and by a combination of both NETs and local teachers ('both') were significantly higher than those taught by local teachers. The effect size in this case was small to moderate at 0.03. This phenomenon will be further analysed in the subsequent analyses.

Table 23. Summary ANOVA table of pre test scores according to teacher mode (Form three)

Teacher	mean score	S.D.
Local	15.31	7.07
Both	16.96	5.13
NET	17.93	5.99
F(2, 456) = 7.00, p = < .01**		
Effect Size = 0.03		

Form Four

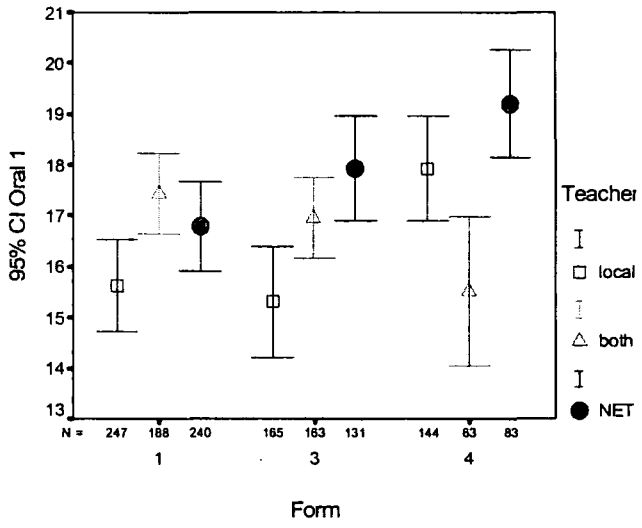
Finally, in the ANOVA analysis carried out on the Form four students, it was found that the scores of students taught by both local teachers and NETs ('both') were on average significantly lower than those taught by NETs only and by local teachers only although the effect size was small to moderate at 0.05. One possible explanation for this result is that at this age when examination preparation takes a high priority, a combined teaching mode is unsettling to students. Full results of the ANOVA are shown on table 24 below.

Table 24. Summary ANOVA table of pre test scores according to teacher mode (Form four)

Teacher	mean score	S.D.
Local	17.92	6.27
Both	15.51	5.88
NET	19.20	4.91
F(2, 287) = 7.27, p = < .01**		
Effect Size = 0.05		

Figure 10 below shows an error bar summary of pre test scores by form and teacher deployment and graphically displays the analysis in the tables above. We can see from the error bars that: in Form one the average score of students taught by 'both' is significantly higher than that taught by local 'local'; in Form three, the average scores of students taught NETs and 'both' are significantly higher than those taught by local teachers, and; in Form four, the average scores of students taught by NETs and locals are significantly higher than those taught by both. One of the sub-groups, Form four 'both' shows a greater variance than the other eight sub-groups.

Figure 10. Error bar summary of pre test scores by form and teacher deployment



ii) Post Test and Teaching Mode

Whole sample

When analysed as a whole sample, there were no significant differences between the mean scores of students taught by the NET teachers and local teachers. However, the

mean scores of students taught by a combination of NET teachers and local teachers were significantly lower than those of students taught by either a NET teacher only or by a local teacher only. There is some evidence therefore that the 'both' teacher deployment is less effective than that of either 'NET' or 'local', although the effect size (eta squared) is small at .003.

The ANOVA summary is given in table 25 below with the effect size as calculated by the eta squared statistic:

Table 25. Summary ANOVA table of post test scores according to teacher mode (whole sample)

Teacher	mean score	S.D.
Local	18.03	6.35
Both	16.84	5.34
NET	18.99	5.89
F (2,913) = 9.81, p n= <0.01**		
Effect Size = 0.003		

In addition, when analysed by age/year group there were also found to be no significant differences between students taught by the three groups of teachers. The results of these one-way analyses of variance are as follows:

Form One

In the case of Form one students, there was no significant difference in the scores of students taught by local teachers and by both local teachers and NETs ('both').

Interestingly however, the mean score of students taught by NETs was higher than the scores of students taught by the other two groups. There is some evidence here that there night be a 'NET effect' at play but this needs to be investigated in subsequent analysis, and it should be noted that the effect size is small at 0.02.

Table 26. Summary ANOVA table of post test scores according to teacher mode (Form one)

Teacher	mean score	S.D.
Local	15.97	6.17
Both	15.87	5.30
NET	17.76	5.91
F (2, 486) = 5.47, p = < 0.01**		
Effect Size = 0.02		

Form Three

We can see from table 27 below that as with the whole sample group there were no significant differences between the mean scores of students taught by the NET teachers and local teachers. Again, following the pattern of the whole sample group the mean scores of pupils taught by either a NET teacher only or by a local teacher only were significantly higher than those of students taught by a combination of NET teachers and local teachers (i.e. 'both'). The effect size was small to moderate at 0.04.

Table 27. Summary ANOVA table of post test scores according to teacher mode (Form three)

Teacher	mean score	S.D.
Local	19.81	6.09
Both	17.93	5.42
NET	20.90	5.72
F (2, 304) = 6.93, p = <0.01**		
Effect Size = 0.04		

Form Four

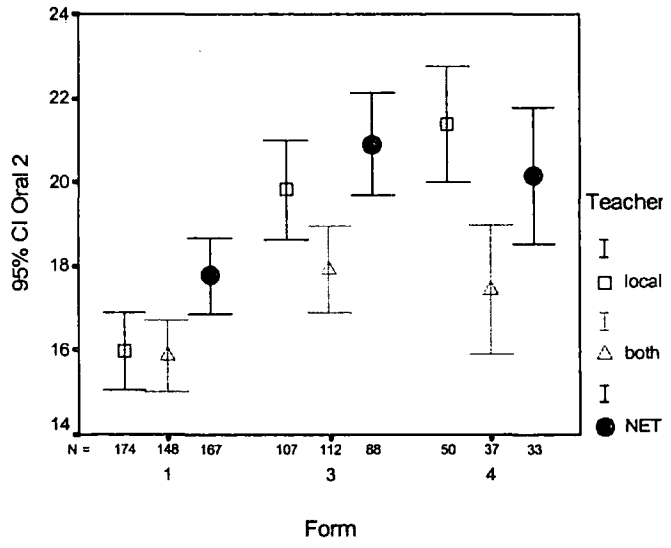
As with the whole sample and the Form three groups we can see from table 28 below that there were no significant differences between the mean scores of students taught by the NET teachers and local teachers. Again however, the mean scores of students taught by *a combination* of NET teachers and local teachers were significantly lower than those taught by either a *NET teacher only* or by a *local teacher only*. In this case, the effect size was large at 0.17.

Table 28. Summary ANOVA table of post test scores according to teacher mode (Form four)

Teacher	mean score	S.D.
Local	21.40	4.84
Both	17.43	4.62
NET	20.15	4.58
F (2, 117) = 7.66, p = <0.01**		
Effect Size = 0.17		

Figure 11 below shows an error bar summary of post test scores according to form and teacher deployment. This figure graphically displays the analysis in the tables above. We can see from the error bars that: in Form one the average score of students taught by NETs is significantly higher than that taught by locals and both; in Form three, the average scores of students taught by NETs and locals are significantly higher than those taught by both, and in Form four, the average scores of students taught by NETs and locals are significantly higher than those taught by both.

Figure 11. Error bar summary of post test scores by form and teacher deployment



6. Teaching Mode and School Level

With regard to the effectiveness of NET teachers, one important question that arises is the extent to which the effectiveness of NETs might vary according to the school level. It is possible, for example, that any NET effect will be different from one school

level to another? Are NET teachers more effective, for example in high level schools? There is currently an on-going discussion in Hong Kong as to whether NET teachers are 'wasted' in lower band schools since those students often do not have a basic level of English language competency and such students are demotivated by NETs since they are not able to follow the teacher's basic instructions. Students in high level schools, it is argued, can maximise the rich input and can make greater language gains especially in listening and spoken English. To explore this question further, a two-way, between groups analysis of variance was conducted to explore any possible interaction between the level of school and the teacher on final (post test) oral assessment scores.

An analysis of the whole sample was firstly undertaken, with teachers divided into three groups: NETs, local and both. A Levene's Test of Equality of Error variances was conducted to verify the underlying assumption of homogeneity of variances, with the following result: $F(8, 907) = 1.829$, $p = 0.37$. The resulting F statistic indicates that such an analysis of interaction is appropriate. The analysis shows that there was a statistically significant main effect for the teacher [$F(2, 907) = 3.42$, $p = 0.03$], however the effect size was very small (eta squared = 0.007). There was also a statistically significant main effect for the school level [$F(2, 907) = 71.78$, $p = <0.01$], with a moderate effect size (eta squared = 0.14). However, the interaction effect [$F(4, 907) = 1.79$, $p = 0.13$] did not reach statistical significance. Post hoc comparisons (using the Tukey HSD test) indicated that there were no significant differences between the mean scores of students taught by NET teachers, local teachers and a combination of both NET and local teachers.

From this analysis of the current data, we can see that there are no significant differences between the mean scores of students taught by these three groups of teachers, and that in addition, there is no interaction between the level of school and the type of teacher deployment. Based on these results, there is little evidence to support the commonly held view that in terms of helping to raise students' listening and speaking skills, NET teachers might be better deployed in high rather than low level schools. As we have seen, there is indeed a significant difference between the mean scores of students taught in high, medium and low band schools but there is no evidence to show that such differences were associated with teacher deployment, i.e. who the students were taught by.

V. SUMMARY FINDINGS

Thus, predictions that one might reasonably make concerning the oral assessment scores seem to be confirmed by this preliminary analysis of the data, namely that:

- there were significant differences between the means of pre and post test scores when the sample was analysed as a whole;
- there were significant differences in the means of pre and post test scores in form three and form four students but not in form one students;
- there was a significant difference between the mean scores of students taught in high, medium and low level schools. These differences were observed when students'

scores were analysed as a whole and separately by age/year. These differences were observed in both pre and post test scores;

- there was a significant difference between the means of students taught in EMI schools and those taught in CMI schools. These differences were observed when students' scores were analysed as a whole and separately by age/year. Additionally, these differences were observed in both pre test and post test scores;
- there were no significant difference between the mean scores of students taught by NET teachers, local teachers and students taught by a combination of both;
- there was no interaction between school level and teacher deployment. This preliminary analysis suggests that the variance in students' scores can be attributed to their school but not to teacher deployment.

This preliminary analysis of frequencies, distributions and comparison of means was carried out using the raw scores obtained from the two administrations of the oral English assessment. A more detailed analysis of this data was conducted using multiple regression modelling and multi-level modelling and is discussed in chapters six and seven. This subsequent modelling however was not conducted using raw scores but using instead Rasch scale scores. The conversion of raw scores to Rash calibrated scores is discussed in the following chapter, together with the theoretical and technical issues involved in Rasch Scale Modelling (RSM).

Footnotes

[1] Recent educational reforms have reorganised the banding system from five bands to three.
[2] Cohen (1988) suggests interpreting the strength of eta squared values on the following guidelines: 0.01 = small effect; 0.06 = moderate effect; and 0.14 = large effect.

CHAPTER 5

RASCH SCALE MODELLING (RSM)

I. CREATING AN OBJECTIVE MEASURE

This chapter discusses the calibration of raw oral assessment scores into Rasch scaled scores in order to obtain independent, objective measure in which the three parameters of item difficulty (δ_i), item steps (τ_j) and person measures (β_n) were obtained. Firstly, the rationale underlying the use of Rasch Scale Modeling (RSM) is discussed, followed by the procedure used for Rasch scores in this study. Finally we will consider the outcomes of this modeling and the resulting implications.

> "When we analyze our data using a Rasch model, we get an estimate of what our construct might be like if we were to create a ruler to measure it. The Rasch model provides us with useful approximations of measures that help us understand the processes underlying the reason why people and items behave in a particular way" (Bond and Fox, 2001:8)

In classical test theory, the difficulty of an item is defined by the proportion of people passing that item (i.e. the item's 'facility index'). Thus an item's difficulty depends directly on the distribution of scores of *others* who have responded to the same item. Bond and Fox (2001) point out the absurdity of "...telling a person that the height of six feet on a ruler depends on what the person is measuring!" (p. 3). Classical test theory in the social sciences thus fails to disentangle the calibration of the item and the measurement of the attribute in question. As a number of writers point out, it is

necessary to construct an objective measure or 'ruler' based not on the raw data but on a calibration of equal intervals between the test respondents and the (difficulty of the) items. In this way, the measuring instrument will be replicable and should better help us to make inferences about the abstraction being measured[1]. RSM is a procedure for transforming raw data into abstract scales of *equal intervals*. In this current research, the pre and post test scores were calibrated into Rasch scores to ensure that estimates of item difficulty[2] were independent of the person measures.

1. First Stage Rasch Scale Modeling

Most researchers calibrating pre and post test raw scores into Rasch scaled scores tend to stop after the initial measure has been constructed, in which the three parameters (item steps, item difficulty and person measures), have been calculated[3]. However, some writers, (e.g. Wolfe and Chui, 1999a, 1999b) have observed that in pretest-posttest measurement of individuals, observed differences over time cannot necessarily be attributed to actual changes in the individual. Although we would expect persons to change from *Time* 1 to *Time* 2 any such observed differences could be the result of either changes in difficulty of individual items over time (i.e. instability) or the different application of the rating scales used in the two sets of measurement. Most analysts tend to assume, without verification, that test items remain stable over time and that the application of rating scales is constant. The Rasch scale modeling conducted in this research follows an equating procedure set out by Wolfe and Chui (1999a) to compensate for any distortions in the measurement of individuals that may arise from changes in how

the same measurement instrument may be applied differently on the two occasions or in how individuals may perceive test items differently on separate occasions.

This procedure uses different anchoring strategies to establish a common frame of reference and to separate the interacting factors involved so that questions concerning the validity of observed changes in time may be adequately addressed. It is argued that through following this procedure, potential misfit to the Rasch Rating Scale Model can be greatly reduced and a more accurate calibrated scale can be obtained. In this thesis, these adjusted, calibrated scores were then used in subsequent, ordinary least squares (OLS) and multi level modeling (MLM) analysis which will be discussed in chapters six and seven.

The calibration of both pre and post test raw oral English assessment scores into Rasch scores was carried out using the WINsteps programme (Linacre, 2001). The procedure employed in the first stage of this calibration followed commonly accepted steps as follows:

- Items with a poor mean square outfit (greater than 1.2 and less than 0.8) were 'deleted'[4] and an anchor file of the remaining items of good fit was created (IAFILE);
- When all items of poor fit had been removed, persons with negative score correlations were subsequently 'deleted'. An anchor file of the remaining persons of good fit was created (PAFILE);
- An item steps anchor file was created (SAFILE);

- The final iteration was run, anchoring on the three files IAFILE, PAFILE and SAFILE;
- The same procedure was carried out for both pre and post test scores.

Separate Time 1 and Time 2 measures were thus established for scale steps (τ_{j1} and τ_{j2}), and items (δ_{i1} and δ_{i2}), which are then compared. However, this stage often reveals problems, since firstly, items are frequently far from the identity line and secondly, the rating scale structure is not common across both times (i.e. it is time dependent). Consequently the meaning of changes in person measures is uncertain and to compensate for this, further analysis is required by proceeding with the next four stages.

II. SECOND STAGE RASCH SCALE MODELING

1. RSM Stage Two Process

The process for establishing a stable frame of reference over time consists of five stages[5] (Wright, 1996a; Wolfe and Chiu, 1999a). Wright (ibid) notes that between two points of time, not only will examinees have changed, but so too will item difficulties, the raters involved and the definitions of the rating scale categories being used. In pretest-posttest measurements, students may for example improve their performance as a result of becoming familiar with the test items being used. Between the two points in time, subjects inevitably mature and change or even withdraw from a particular programme altogether. There is also a much-reported regression towards the mean, which further

confounds any given measures. Wolfe and Chiu (1999a) point out that even though such changes might be small, "...[they] present potential confounds [which] may distort the measurement of change, making it unclear whether the observed changes in the outcome variable are due to the intervention or some other effect" (p135). In order to make valid comparisons between pretest and posttest measurements, a stable frame of reference is required so that in comparing performance over time we can either eliminate or control any other changes that may have occurred. Wright (1996a) proposes an equating procedure that enables the researcher to disentangle these potential confounds and create a stable frame of reference so that "the functioning of test items and rating scales remain constant across time" (p478). The procedure for corrected Time 1 to Time 2 comparisons is fully described by Wright (1996a) and has been subsequently illustrated by Wolfe and Chiu (1999a) in a five-stage algorithm. This procedure was the one used in the current study.

By producing separate Time 1 and Time 2 calibrations, RSM permits us to evaluate whether the item calibrations (δi) are stable across samples of persons, and similarly whether person measures (β_n) are stable across samples of items. Wolfe and Chiu (1999a, 1999b) refer to this as 'invariance evaluation', and from the estimates produced in the first stage of the RSM process can be computed using equation 5.1 below. In this computation, the stability of the two parameters ($\hat{\theta}_1 \; and \; \hat{\theta}_2$) obtained on two separate occasions, are evaluated by computing the standardized differences between the two occasions (Time 1 and Time 2). Those standardized differences that conform to RSM would be expected to have a value of 0.00 and a standard deviation of 1.00. Estimates

with values greater than ±2.00 indicate less stability over the two points in time than would be expected and would be treated accordingly (see section 2 below)[6].

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{[SE(\hat{\theta}_1)]^2 + [SE(\hat{\theta}_2)]^2}} \qquad (5.1)$$

2. Evaluating Change over Time

In the second stage, the data was 'stacked' vertically and each person is treated as unique at each point of time, appearing twice (Time 1 and Time 2). The purpose of this stage is to create common scale calibrations (τ_{jc}), which are then used as anchors to obtain corrected measures in the subsequent stages. This is the common frame of reference referred to by Wright (1996a). In this stage, items that displayed poor fit in stage 1 often show greater misfit confirming that the items do indeed function differently from time 1 to Time 2.

In the third stage, corrected person measures (β_{n1c}) and item calibrations (δ_{i1c}) for Time 1 were obtained by anchoring on the rating scale steps (τ_{jc}) produced in Stage 2.

In the fourth stage, the same common rating scale steps (τ_{jc}) produced in Stage 2 were used as anchors to obtain corrected person measures (β_{n2c}) for Time 2. Items from Stage 1 that were found to be stable over time were anchored on (δ_{i1c}).

In the fifth and final stage, corrected item calibrations (δ_{i2c}) were obtained for Time 2 data, with the rating scale anchored on τ_{xc} and the persons anchored on β_{n2c}. It was then possible to analyse the change in item difficulty by computing $\delta_{i1c} - \delta_{i1c}$.

III. RESULTS OF RSM

The results of the RSM, both before and after the adjustment procedure was applied are outlined in the sections below.

1. Uncorrected and Corrected item measures

It was found that the item fit was significantly improved by following this correction procedure. Table 29 below shows the corrected and uncorrected item measures from time one and time two. The relative item difficulty remained the same with grammar being the most 'difficult', confirming anecdotal evidence that teachers tend to be less tolerant towards grammatical errors than they are to other aspects of oral production[7]. Alternatively, students could indeed be 'worse' at grammatical form than they are, for example, at comprehension[8].

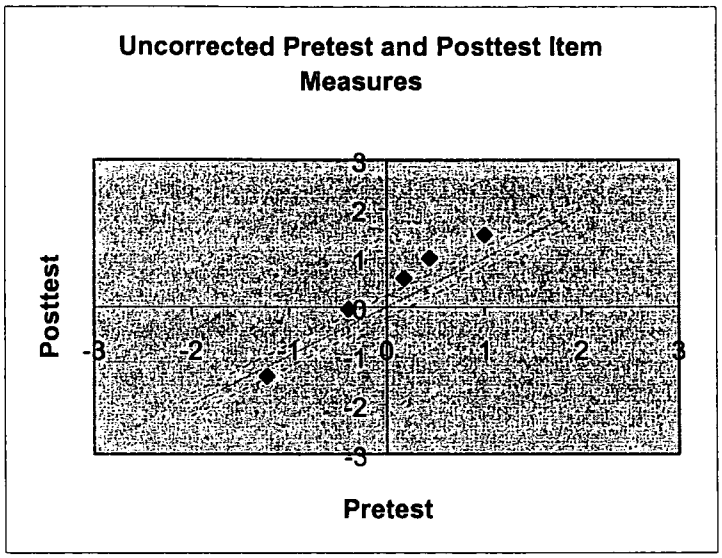
Table 29. Comparison of time one and time two corrected and uncorrected item measures

Item	Uncorrected Time 1 measures (δ_{i1})	Uncorrected Time 2 measures (δ_{i2})	Corrected Time 1 measures (δ_{i1c})	Corrected Time 2 measures (δ_{i2c})
grammar	1	1.47	0.96	0.81
vocabulary	0.44	1	0.41	0.62
fluency	0.18	0.58	0.19	0.27
pronunciation	-0.39	-0.05	-0.36	-0.33
comprehension	-1.23	-1.42	-1.19	-1.15

NOTE: Lower scores indicate easier items

The scatterplot in figure 12 below[9], shows that before carrying out this correction procedure, only one item, comprehension, was within the 95% confidence intervals (indicated by the dotted horizontal lines), whilst the other four were noticeably outside the confidence intervals.

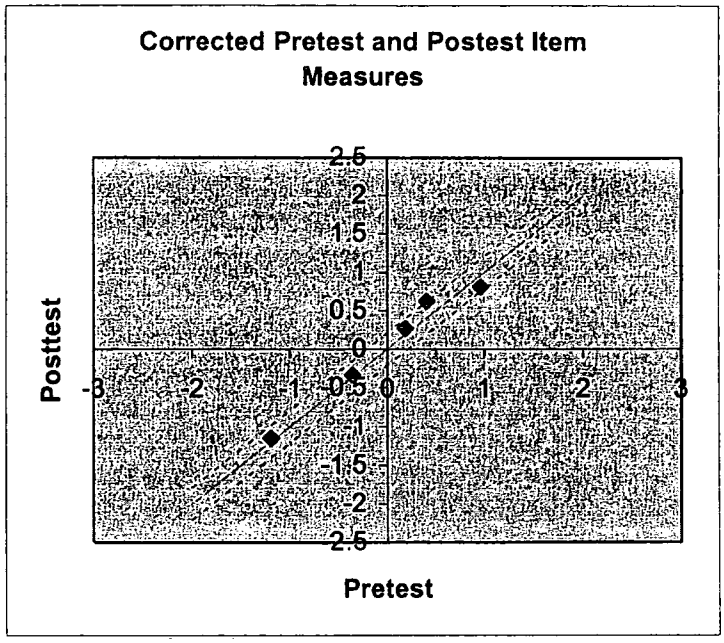
Figure 12. Scatterplot of uncorrected pre and post tests item calibrations



Note: The unbroken diagonal line in the center is the pretest-posttest identity line, while the two parallel dotted lines on either side represent the 95% confidence intervals.

However, upon completion of Wright's (1996a) correction procedure, all five items remained within the confidence intervals, as shown in Figure 13 below, thus adding more evidence to support the view that through this correction procedure there is less error in the resulting calibration.

Figure 13. Scatterplot of corrected Time 1 and Time 2 item calibrations



Note: The horizontal line in the center is the pretest-posttest item fit line, while the two parallel dotted lines on either side represent the 95% confidence intervals.

2. Uncorrected and Corrected Student Measures

In addition to obtaining corrected item measures, the procedure adopted in this study also produced corrected measures for students. Figure 14 below shows the uncorrected and corrected measures for the students to whom the oral assessment was administered on both occasions. Whilst the study completed by Wolfe and Chiu (1999a) showed considerable differences between the uncorrected and the corrected standardized differences[10], in this analysis little systematic difference was evident on the scatter plot between the uncorrected and the corrected measures. This is illustrated by the fact that the number of plotted points appears to be roughly the same on both sides of the identity line. We cannot therefore say that the students' standardized differences were greater or less according to whether the uncorrected or the corrected measures are considered.

However, there are some noticeable differences between the two sets of measures, particularly for those students who had negative standardized differences (i.e. those who were more sensitive to the test items between the pretest and posttest). The correlation coefficient (r = 0.86) is reasonably high but the dispersal of plots in the lower left quadrant of Figure 14 appear to be slightly wider, as illustrated by group A. There are examples of large individual differences in the uncorrected and corrected differences of students' standardized differences. The student represented by point B in Figure 14 had the largest discrepancy between the uncorrected and the corrected student measures of 9.07 (i.e. $z_{uncorrected} = 5.22$ and a $z_{corrected} = -3.85$).

Figure 14. Scatter plot of uncorrected and corrected student standardized differences

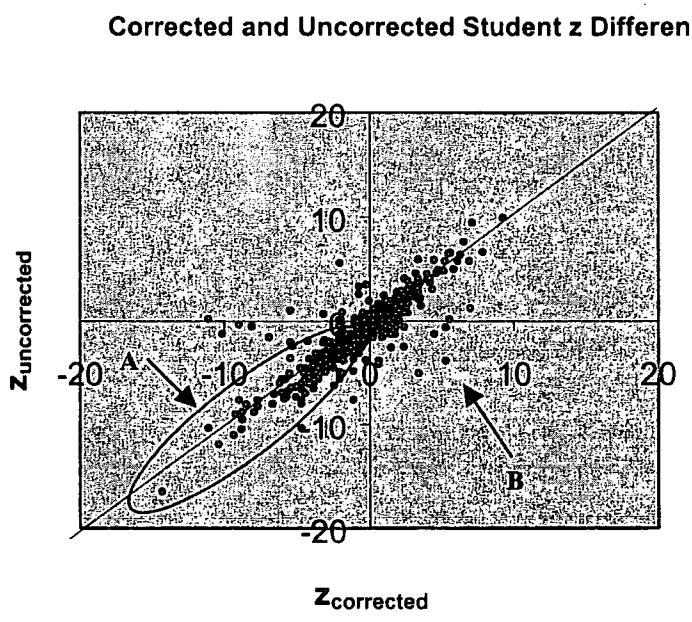


Table 30 below gives a summary of corrected and uncorrected student measures. We can see that firstly, following this correction, there were fewer misfitting students (10.6% uncorrected, 9.3% corrected). Secondly, as a result of the correction, there were

fewer students with absolute significant standardized differences greater than two (13.6% uncorrected, 10.8% corrected). On the other hand however, in contrast to Wolfe and Chiu's (1999a) findings, mean pretest and posttest measures in this study decreased slightly as can be seen by the mean standardized differences (Mean $z_{uncorrected}$ = -0.72, mean $z_{corrected}$ = -1.31). The difference between the students' prestest and the posttest measures would therefore depend on whether the uncorrected or the corrected measures are used. Finally, not only would our interpretation of the amount of change be different depending on whether the corrected or the uncorrected measures are used, but so too would our interpretation of which students had changed between the pretest and the posttest. If our assumptions were based on uncorrected standardized differences in the measures as opposed to the corrected measures, we would not draw the same conclusion about how the students had changed over time[11].

Table 30. Uncorrected and Corrected Student Measure Summary Statistics

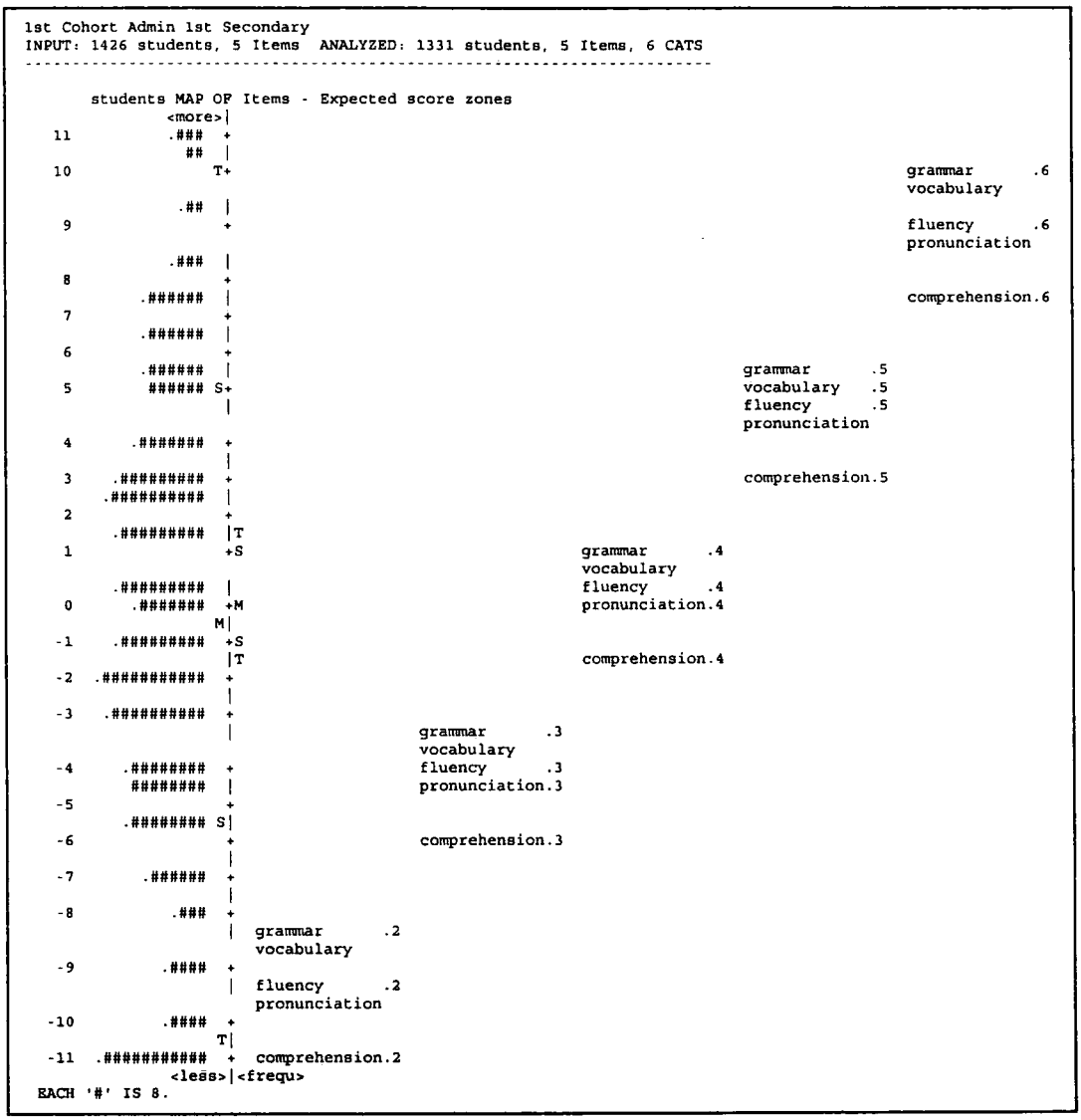
Statistic	Uncorrected Measures	Corrected Measures
Number with Fit > 2.00	153 (10.62%)	134 (9.3%)
Mean (z)	-0.72	-1.31
Number with Significant z	98 (13.6%)	78 (10.83%)
SD (z)	2.79	2.94

Number with Fit statistic represents the number of items with fit statistics > 2.00 summed for both occasions and the percentage therefore is the total number of misfitting items divided by 10 (5 test items x 2). Mean (z) is the mean standardized difference of all 5 items, and SD (z) is the standard deviation of the standardized differences. The number with Significant z is the number of items with absolute standardized differences >2.

3. Holistic Versus Discrete Criteria

One of the difficulties that arose during the RSM in this study centred on the finding that teachers were on the whole unable to distinguish between the different components of the criteria and tended to mark the students holistically[12]. This finding is supported by the raw Time one scores for the five assessment items (fluency, comprehension, grammar, pronunciation and vocabulary) in which correlations ranged from .82 to .89 in the pre test and .80 to .88 in the post test (see Tables 13 and 14, p. 90). Figure 15 below, illustrates the lack of separation in the calibrations since the five different items marked by the assessors are all closely grouped. We can see for example, that if an assessor rated a student six in grammar, he/she was likely to award the same mark for vocabulary (or indeed for fluency, comprehension or vocabulary resource). Whilst the distribution of students on the left side of the ruler seems more or less normal, that of the items on the right side does not.

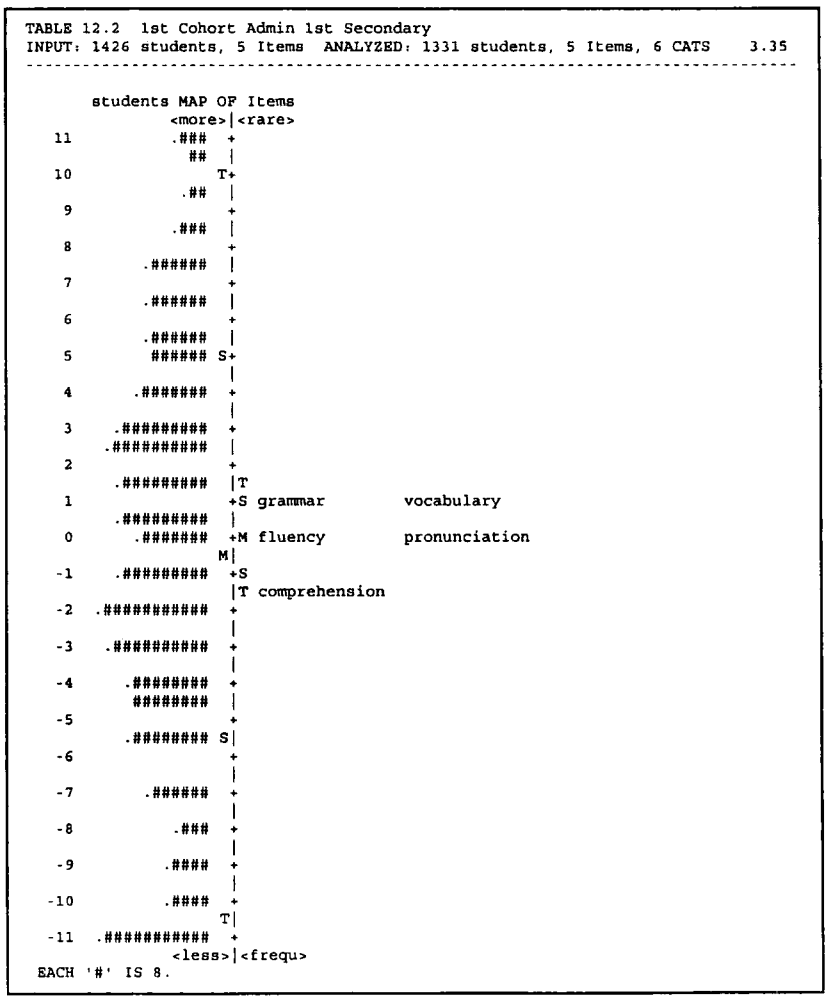
Figure 15. Item map showing expected score zones of calibrated measure of students and items



Further evidence of the holistic nature of the assessors' marking can be seen in figure 16 below, in which the overall item map shows the five different criteria grouped very much

together, with little discrimination between them. One might have expect that RSM would have produced a result in which individual items are more clearly separated, with some being more difficult than others. In English language testing, this is likely to occur in a multiple choice reading test, for example, in which the test items display a range of difficulty, with each one having a distinctly different facility index. This was not the case in this study.

Figure 16. Overall item map of calibrated student and item measures



Another finding that possibly adds weight to the view that assessors were unable to apply the criteria discretely lies in the results of correlations between the raw scores and the Rasch measures that resulted from the adjustment procedure outlined above[13]. As we can see from Table 31 below, the correlations between the raw scores and the respective Rasch calibrated scores are noticeably high. This tends to suggest that, at least in terms of the item measures, the RSM process employed in this study did not produce discrete calibrations that were noticeably different from the raw scores[14]. It is felt however, that this is not the result of the calibration procedure itself, which did indeed produce some technical improvements to the calibrated scores (e.g. better fit of item and student), but was rather a characteristic of the way in which assessors marked their students.

Table 31. Correlations between Time 1 and Time 2 raw scores and Rasch calibrated measures

	Time I RSM measure	Time 2 RSM measure	Time 1 raw score
Time 1 raw score	0.99	0.67	
Time 2 raw score	0.67	0.99	0.66

Note: Time 1, n = 1424; Time 2 n, = 916

Despite the limitations of the RSM alluded to above, the final Time one and Time two Rasch calibrated student measures (adjusted) were used in the subsequent modeling of the data. The next chapter continues the data analysis by conducting multiple regression analysis on the adjusted Rasch scores that have just been outlined in this chapter.

Footnotes

[1] For a more complete theoretical and technical rationale, see Wright and Mok (2000).

[2] In a more traditional paper and pencil test, the test items would refer to the individual questions that respondents answer. In this context, 'item' and 'item difficulty' refer to the different components of the oral assessment i.e. grammar, vocabulary, fluency, pronunciation and comprehension and their relative difficulty.

[3] Rasch (1960) noted that in his models person and item parameters were fully separable – a property he referred to as 'specific objectivity'.

[4] A 'deleted' item in this case means that it will not be used for subsequent anchoring i.e. deleted items are only *temporarily* removed from the modeling process.

[5] Wolfe and Chiu (1999a, 1996b) refer to using five 'steps' to carry out this equating procedure. This thesis will use the term 'stages' to avoid confusion when referring to *item* steps.

[6] See Wright and Masters (1982).

[7] In Barratt and Kontra's (2000) study, students commented that "...[Native-speaking] NS teachers were less sharp in presenting grammar and less vigilant in correcting errors than their host colleagues" (p.21).

[8] This could also be a result of classroom practice in which traditionally in English classes in Hong Kong students are given little opportunity to practice the target language in meaningful contexts, while the teacher tends to talk a lot in a more teacher-centred style of classroom interaction.

[9] With thanks and acknowledgements for help with this and the following figure to Ms Fung Suk-yee, Tammy (CRIC, HKIEd).

[10] In the Wolfe and Chiu (1999a) study, the standardized differences based on corrected teacher measures were greater than those based on uncorrected measures, as evidenced by the fact that the majority of the plots fell above the identity line.

[11] This may however be due to the 'deletion' of cases/items in the RSM process itself.

[12] This is not necessarily considered by all writers to be a negative factor. Oller (1976) for example, stated that "...it is my opinion that so-called integrative tests are better than discrete point tests." (p. 161)

[13] This is arguably a moot point however, since high correlations can also occur between raw scores and well measured calibrated scales. This could also be a manifestation therefore of the close relationship between RSM and classical test theory.

[14] It will be remembered that the correlations between the sub-components of the raw scores were also very high [see Chapter 4], with coefficients ranging from 0.82 to 0.89.

CHAPTER 6

REGRESSION ANALYSIS

1. THE REGRESSION MODEL

This chapter proceeds with the data analysis described in chapters four and five. It also aims to fit a model to the data from the oral assessment and subsequently use that model to predict values of the dependent variable(s) from one or more independent variables[1]. This linear model, is represented in equation 6.1 below, in which: Y is the outcome variable, β_0 is the constant, β_1 is the coefficient of the first predictor (X_1), β_2 is the coefficient of the second predictor (X_2), β_n is the coefficient of the nth predictor (X_n) and ε_i is the difference between the predicted and the observed value of Y for the ith student (i.e. the residuals or error).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \varepsilon_i \qquad (6.1)$$

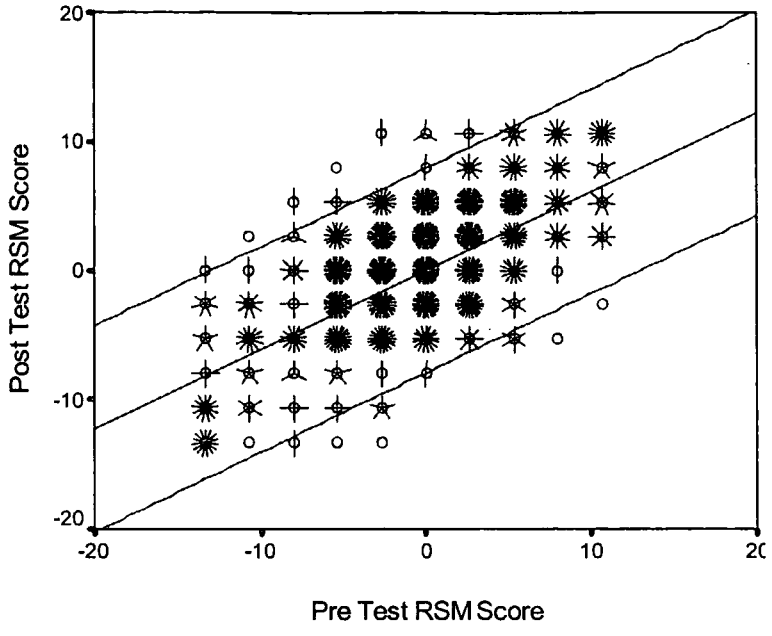
The primary aim of this stage of the analysis was to build a model that would explore and investigate any possible 'NET effect' on students' English language oral proficiency and, in addition, to explain or account for as much of the variation in the model as possible, as represented by the adjusted R^2 statistic[2]. It was hoped that the model would be able to predict as accurately as possible the students' outcome variable from given predictor variables. In this study, the outcome variable Y is the *time two* students' oral assessment score to be predicted from a number of predictor variables built into the model (e.g. *time one* students' oral assessment score, school level, medium of instruction, etc. [see Chapter 4]). This analysis was carried out using the statistical software package SPSS 11.0 (SPSS Inc, 2001).

1. Basic Correlation

The basic regression model for the whole sample (i.e. combined form 1, form 3 and form 4) in this study was built from the Time one and Time two Rasch calibrated oral assessment scores [see Chapter 5 and Appendix II], illustrated in Figure 17 below, in which r = 0.67 (p = <0.01, n = 916). The resulting R^2 statistic of 0.45 thus represents 45.3%[3] of the variation. The multiple regression analysis which follows attempted to build a more accurate model in which more of the variance could be explained by the inclusion of more predictor variables whose β coefficient values would be statistically significant and make a positive contribution to the final regression equation(s) 6.1, above.

The starting point for this analysis is to determine whether or not this type of analysis is suitable for the type of data generated from this research. The correlation between Time one and Time two (Rasch calibrated) oral assessment scores described above, and illustrated in Figure 17 below suggests a good estimate of the overall fit of the basic model and indicates that the Time one oral assessment score is a good predictor variable of the outcome variable, Time one (RSM) score.

Figure 17. Scatterplot showing relationship between Time 1 and Time 2 Rasch scores for oral English assessment (whole sample)



Note: R squared value = .43. The central diagonal line represents the fit line and the two parallel lines on either side represent the 95% confidence intervals.

II. CHECKING ASSUMPTIONS

1. Underlying Assumptions

A number of writers (e.g. Tabachnick & Fidell, 1996; Field, 2000) emphasise the importance of satisfying various assumptions when conducting multiple regression analysis. These assumptions are discussed in more detail by Berry (1993) and the extent to which they have been satisfied in this analysis is briefly considered below.

i) Variable types

In this study, all the potential predictor variables (see Chapter 4, I.2) satisfied the assumption that they were ordinal or categorical and the outcome variable was

quantitative and continuous. Dummy variables were created for medium of instruction, school level, student level (as established by the students' regular English teacher) and for NET teacher.

ii) Sample size

It is felt that the sample sizes used in this study were suitable although due to the parameters established by MENETS, the form four sample size was relatively smaller than the others and could ideally have been larger. The sample sizes were as follows:

whole sample:	n = 790
Form one:	n = 420
Form three:	n = 250
Form four:	n = 120

iii) Multicollinearity and singularity

According to this assumption, there should be no perfect linear relationship between the predictive variables i.e. they should not correlate too highly since this would lead to problems of interpretation. Generally predictor variables did not correlate highly with one another in this analysis. In the 'whole sample analysis', correlations ranged from .62 to -.50[4] and the relationships between the predictor variables and the outcome variable (RSM2) were significant but not too high. The exception to this finding was in the Form four analysis in which the correlation between 'medium of instruction' (MOI) and 'school level' (d_schlv) was 1.00. In this instance, the sample size was small (n = 120) and can probably be attributed to the fact that in this instance all the high level schools (band one) were EMI and the lower level schools were all CMI[5].

iv) Normality and normally distributed errors

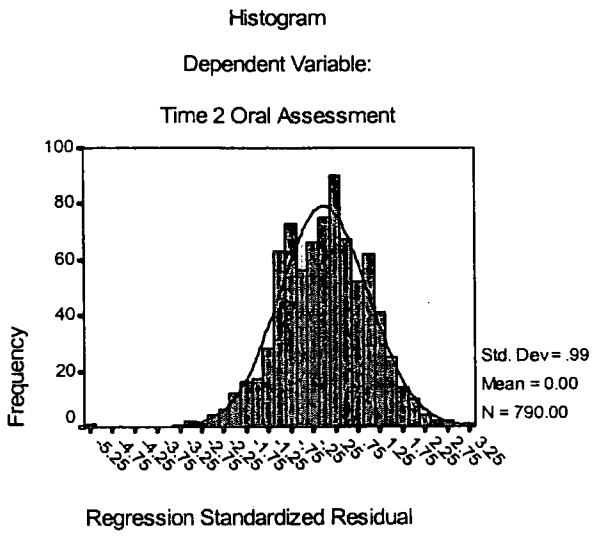
The residuals in regression models are assumed to be random and normally distributed with a mean of zero. It would be expected therefore, that 95% of the cases would have standardised residuals within plus or minus two. An analysis of basic Time One and Time two regression residuals for each of the four sample groups produced the following results:

Table 32. Summary table of casewise diagnostics

Sample group	sample n	n ±2	% ±2
whole sample	790	33	4.18
Form one	420	17	4.08
Form three	250	11	4.40
Form four	120	6	5.00

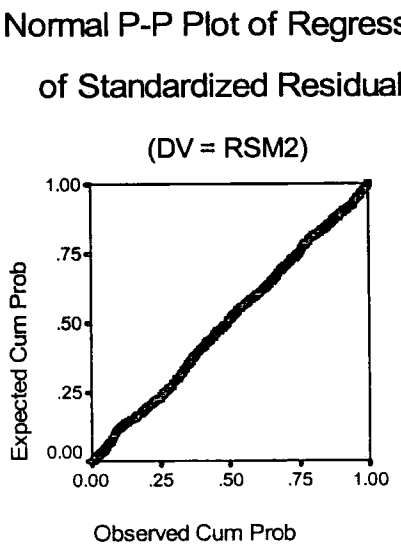
All of the sample groups show standardised residuals equal to or less than five per cent, in conformity with what one might expect. In addition, an analysis was carried out to determine the normality of the distribution of the residuals[6]. As we can see from figure 18 below, the distribution is more or less normal, with only a slight skew, thus giving us more confidence in the basic model.

Figure 18. Histogram showing distribution of standardised residuals (whole sample)



As regards deviations from normality of the residuals, figure 19 below shows that the points hardly vary from the straight diagonal line (which represents a normal distribution). Similar analyses of the distributions in the other samples (form one, form three and from four) were also carried out, with similar results.

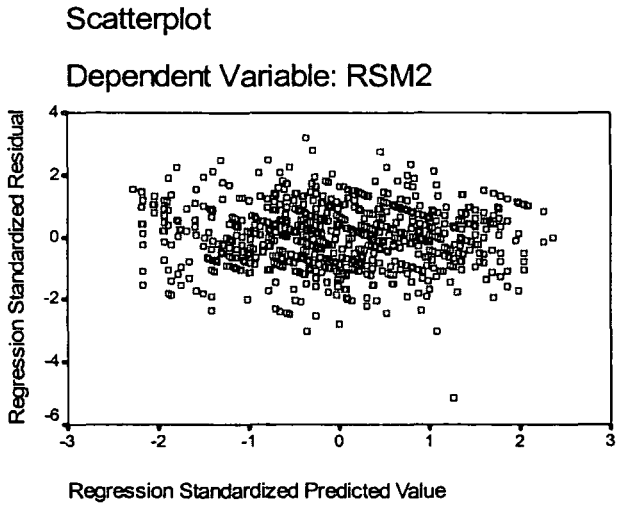
Figure 19. Scatterplot showing normal probability plot of standardised residuals (whole sample)



v) Homoscedasicity

The scatterplot in figure 20, below shows how the points are more or less randomly dispersed throughout the plot, suggesting that the assumptions of homoscedasicity and linearity have been met. If the scatterplot had shown any clustering or linear pattern, we might question whether or not at each level of the predictor(s) the residuals have the same variance.

Figure 20. Scatterplot showing regression of standardised residuals and standardised predicted values



2. Choosing the Model

Researchers are also cautioned against simply entering all possible variables into a model and 'hoping for the best' and subsequently constructing a theory based on the outcome. Tabachnick and Fidell (1996) note that researchers usually assign the order of entry of variables according to 'logical or theoretical considerations' (p. 149). In this study, a *hierarchical* linear regression[7] was adopted since the earlier analysis of means, plus a theoretical rationale based on second language acquisition and the Hong Kong teaching-learning context (see Chapter 4). Thus for example, the predictor

variable 'medium of instruction' was considered to be crucial in predicting the outcome variable, since students in EMI secondary schools in Hong Kong tend to have far more exposure to the English language during regular (non-English language subject) lessons that are taught in the English medium than they do during English language lessons. Through their greater exposure to English, EMI students are thus likely to make greater gains than their CMI counterparts. Similarly, the literature on 'self-fulfilling prophecies' (e.g. Rosenthal, 1990, Downey et al 1996, Hurley, 1997) suggests that students in a high band school are also likely to make greater gains than their peers in low band schools[8] due to expectations from teachers, parents, schools and students themselves. A further variable included in this modelling include the teacher's assessment of their students as being either 'high', 'medium' or 'low' ability[9]. Finally of course the model included the variable of whether the students were taught by a NET teacher, a local teacher or some combination of both. In summary, the predictor variables included in this modelling were:

- Time 1 oral assessment;
- Dummy variables for school level;
- Medium of instruction;
- Dummy variables for students' ability in the two years of the study;
- Dummy variable indicating whether students had been taught mostly by a NET teacher[10], or some combination of a NET teacher and a local teacher in the two years of the study.

III. REGRESSION ANALYSIS RESULTS

1. Whole Sample

In the first analysis, all of the students' Time one and Time two results were modelled together as a 'whole sample'. The predictor variables were added to the model in five stages, with the results shown in table 33 below.

Table 33. Summary of hierarchical regression models for outcome variable (RSM2) from predictor variables (Whole Sample, n=790)

Variable	*B*	*SE B*	β
Model 1			
Time one oral assessment	0.57	0.02	0.62**
Model 2			
Time one oral assessment	0.49	0.03	0.54**
English Medium of Instruction	2.07	0.30	0.20**
Model 3			
Time one oral assessment	0.49	0.03	0.53**
English Medium of Instruction	3.12	0.54	0.31**
School level (high)	-0.62	0.62	-0.06
School level (medium)	0.99	0.38	0.09**
Model 4			
Time one oral assessment	0.33	0.03	0.35**
English Medium of Instruction	3.18	0.50	0.32**
School level (high)	-0.24	0.57	-0.02
School level (medium)	1.08	0.35	0.10**
Student level (year 1) high	-0.47	0.45	-0.04
Student level (year 1) medium	-0.02	0.38	0.00
Student level (year 2) high	4.87	0.48	0.41**
Student level (year 1) medium	2.56	0.37	0.25**
Model 5			
Time one oral assessment	0.32	0.03	0.34**
English Medium of Instruction	2.93	0.50	0.29**
School level (high)	0.03	0.57	0.00
School level (medium)	1.04	0.34	0.10**
Student level (year 1) high	-0.39	0.45	-0.04
Student level (year 1) medium	-0.07	0.38	0.00
Student level (year 2) high	4.88	0.47	0.41**
Student level (year 2) medium	2.56	0.37	0.26**
NET teacher (year 1)	1.00	0.28	0.09**
NET teacher (year 2)	0.57	0.57	0.03

*p < .05 **p = < .01

In the final model five shown in table 33 above, six of the ten β coefficients were significant, albeit with relatively small values with student level (year two), '*time one* assessment scores' and 'medium of instruction' producing relatively high β coefficients of .41**, .34** and .29** respectively. Although the predictor variable 'NET teacher (year one)' produced a significant β coefficient (.09**) the value is relatively small suggesting that the predicted NET effect in students' oral English proficiency gain is very small.

Table 34 below gives a summary of all the related R values for the five models constructed in this whole sample analysis. It is felt that the final multiple correlation coefficient of R = .73** (R^2 = .53, Adjusted R^2 = .52) was relatively high, with the adjusted R^2 statistic showing that some 52% of the variation was explained by the predictor variables. (A full table of partial correlations is shown in Appendix IV)

Table 34. Sequential regression model summary showing relative R values

Model	R	R^2	Adj R^2	R^2 Change
1	.62	.38	.38	.38
2	.65	.42	.42	.04
3	.65	.43	.42	.01
4	.72	.52	.52	.09
5	.73	.53	.52	.01

After the first step one, with the time one oral assessment (RSM1) added to the equation, R^2 = .38, F_{inc} (1, 788) = 486.35, p < .01**. After step two, with medium of instruction added to the equation, R^2 = .42, F_{inc} (1, 787) = 48.06, p < .01**. After step three, with high and medium school level added, R^2 = .42, F_{inc} (2, 785) = 6.85, p < .01**. After step four, with high and medium student levels added for both years one

and two, $R^2 = .52$, F_{inc} (4, 781) = 38.18, $p < .01$**. Finally, after step five, with NET teacher added for both years one and two, $R^2 = .53$, F_{inc} (2, 779) = 8.69, $p < .01$**.

The final step five model summary reveals that with all the predictor variables in the equation, R = .73, F (10, 779) = 88.34, $p < .01$**. It is felt that this final whole sample model was able to satisfactorily explain a high percentage of the gain between Time one and Time two.

2. Form One

Following the analysis of the whole sample, the modelling then proceeded to investigate the students by age/class. Of the total sample, 420 students were from Form one classes. Firstly, again in order to establish a firm basis on which to carry out the regression analysis, the correlations between the outcome variable and the predictor variables were calculated, producing *r* values ranging from .62 to .15. A full correlation matrix for Form one students is shown in Table 35 below.

Table 35. Intercorrelations between dependent variable (RSM2) and predictor variables (Form one)

	RSM2	RSM1	MOI	d_sch_lv	d_stu_lv
RSM1	.62				
MOI	.45	.38			
d_sch_lv	.40	.36	.93		
d_stu_lv	.42	.45	.09	.06	
d_net	.15	.03	.03	.25	-.01

A hierarchical multiple regression was conducted on this group of students in a similar fashion to the whole sample, although important changes were made to the

dummy variables. Firstly, the predictor variables 'school level (high)' and 'school level (medium)' were combined into one, new dummy variable[11] although the correlation between this new variable and MOI remained high at .94. Secondly, the teachers' rating of their students in years one and two[12] was also rationalised from the four dummy variables used in the whole sample study (see table 35 above) into just one new dummy variable, 'student level'[13]. A summary of the unstandardised coefficients (*B*), standard errors (*SE B*) and standardised coefficients (β) is shown in Table 34 below.

Table 36. Summary of hierarchical regression models for outcome variable (RSM2) from predictor variables (Form one, n = 420)

Variable	*B*	*SE B*	β
Model 1			
Time one oral assessment	.55	.03	.62**
Model 2			
Time one oral assessment	.47	.03	.52**
English Medium of Instruction	2.51	.39	.25**
Model 3			
Time one oral assessment	.47	.03	.53**
English Medium of Instruction	4.51	.97	.46**
School level (high and medium)	-2.15	.95	-.22*
Model 4			
Time one oral assessment	.38	.04	.43**
English Medium of Instruction	4.29	.95	.44**
School level (high and medium)	-1.70	.93	-1.75
Student level	1.93	.39	.20**
Model 5			
Time one oral assessment	.39	.04	.44**
English Medium of Instruction	4.70	.95	.48**
School level (high and medium)	-2.39	.95	-.25*
Student level	1.90	.39	.19**
NET teacher	1.10	.35	.11**

*p < .05 **p = < .01

As we can see from the coefficients in Table 36 above, all of the values were significant at the p = <.001 level, with Time one assessment and MOI producing relatively high values (.44** and .48** respectively). Although the other three predictor variables in the model (school level, student level and NET teacher) are also

statistically significant, their values are relatively low. This suggests that the Time one oral assessment and the medium of instruction are relatively strong predictors of the outcome variable, school level, student level and NET teacher are only weak (yet statistically significant) predictors.

Secondly, the relative statistics for the multiple correlation coefficient (R), R Square (R^2), the adjusted R Square (Ad R^2) and the R^2 were also calculated and are reported in Table 37, below.

Table 37. Sequential regression model summary showing relative R values (Form one)

Model	R	R^2	Ad R^2	R^2 Change
1	.62	.39	.39	.39
2	.67	.44	.44	.06
3	.67	.45	.45	.01
4	.69	.47	.47	.03
5	.70	.48	.48	.01

Table 37 above shows that through this modeling, the percentage of variation increased from 39% in model one to some 48% in the final model five. It can be observed from the form one modeling, that after the first step one, with the time one oral assessment (RSM1) added to the equation, $R^2 = .39$, F_{inc} (1, 418) = 264.21, $p < .01$**. After step two, with medium of instruction added to the equation, $R^2 = .44$, F_{inc} (1, 417) = 41.74, $p < .01$**. After step three, with a dummy variable for school level added, $R^2 = .45$, F_{inc} (1, 416) = 4.98, $p < .01$**. After step four, with a dummy variable for student level added (for years one and two), $R^2 = .48$, F_{inc} (1, 415) = 23.74, $p < .01$**. Finally, after step five, with a dummy variable for NET teacher added (for years one and two), $R^2 = .49$, F_{inc} (1, 414) = 9.76, $p < .01$**. The final step five model with all the predictor variables in the equation , R = .70, F (5, 414) = 80.00, $p < .01$**.

It is reasonable to suggest that although the standardized coefficient values (β) of student level and NET teacher were relatively low, they nevertheless did add collectively to the prediction above and beyond what could be predicted from Time one oral assessment and MOI alone.

Form One Summary Equations

Finally, from the modeling of the form one students, we can formulate a least-squares regression (prediction) equation for the outcome variable *Time two* oral assessment. Equation 6.2 below shows this equation using unstandardised 'Beta' values:

$$\textit{Predicted RSM2} = -2.50 + (.39)*(\textit{RSM1}) + (4.70)*(\textit{MOI}) - (2.39)*(\textit{D\_STU\_LV}) + (1.90)*(\textit{D\_STU\_LV}) + (1.10)*(\textit{D\_NET}) \quad (6.2)$$

This same equation may be expressed with all of the values converted into standardized Z-scores, as shown in equation 6.3 below:

$$\textit{Predicted } Z_{RSM2} = 0 + (.44)*(Z_{RSM1}) + (.48)*(Z_{MOI}) - (.25)*(Z_{D\_SCHLV}) + (.19)*(Z_{D\_STU\_LV}) + (.11)*(Z_{D\_NET}) \quad (6.3)$$

This analysis then proceeded to construct similar models for form three and from four students.

3. Form Three

The respective correlations between the outcome variable and the predictor variables were first of all calculated, with Form three students producing *r* values ranging from .62 (RSM1) to -.08 (NET). A full correlation matrix for Form three students is shown in Table 38 below.

Table 38. Intercorrelations between outcome variable (RSM2) and predictor variables (Form 3)

IVs	RSM2 (DV)	RSM1	MOI	d_sch_lv	d_stu_lv
RSM1	.62				
MOI	.40	.41			
d_sch_lv	.34	.46	.69		
d_stu_lv	.37	.39	.10	.10	
d_net	-.08	-.13	-.14	.00	-.09

We can see from table 38 above that the correlations between the outcome variable and the predictor variables for Form three students were very similar to the From one correlations – in fact the correlation coefficient *r* of .62 for RSM2 and RSM1 was identical. For Net teacher however, the coefficient was much lower at .08, p = .09 (not sig.).

A summary of the unstandardised Beta coefficients (*B*), standard errors (*SE B*) and standardised coefficients (β) is shown in table 39 below.

Table 39. Summary of hierarchical regression models for outcome variable (RSM2) and predictor variables (Form three)

Variable	*B*	*SE B*	β
Model 1			
Time one oral assessment	.58	.05	.62**
Model 2			
Time one oral assessment	.52	.05	.55**
English Medium of Instruction	1.84	.56	.17**
Model 3			
Time one oral assessment	.53	.05	.57**
English Medium of Instruction	2.32	.72	.22**
School level (high and medium)	-.80	.74	-.07
Model 4			
Time one oral assessment	.47	.05	.50**
English Medium of Instruction	2.36	.71	.23**
School level (high and medium)	-.67	.73	-.06
Student level	1.75	.59	.16**
Model 5			
Time one oral assessment	.47	.05	.50**
English Medium of Instruction	2.43	.72	.23**
School level (high and medium)	-.73	.74	-.07
Student level	1.77	.59	.16**
NET teacher	.27	.52	.03

*p < .05 **p = < .01

As we can see from table 39 above, in the final model, three of the five predictor variables produced significant standardised Beta coefficients (β), with two of the values for RSM1 and EMI being relatively high (.50** and .23** respectively). Of the other two predictor variables (school level and NET teacher), the β values were low and not statistically significant as we can see from the high Standard Errors relative to the unstandardised Beta coefficients. This suggests that for Form three students, the school level and NET teacher were not good predictor variables for the regression model. To substantiate this, the relative statistics for the multiple correlation coefficient (R), R Square (R^2), the adjusted R Square (Ad R^2) and the R^2 were also calculated and are shown in table 40, below.

Table 40. Sequential regression model summary showing relative R values (Form three)

Model	R	R^2	Adj R^2	R^2 Change
1	.62	.39	.39	.39
2	.64	.42	.41	.02
3	.65	.42	.41	.00
4	.66	.44	.44	.02
5	.66	.44	.44	.00

Table 40 above shows that the percentage of variation as shown by the adjusted R^2 statistic, increased from 39% in model one (the same percentage as form one students), to 44% in the final model five (compared to 48% in form one students). The inclusion of the variables 'school level' in model three and 'NET teacher' in model five did not increase the R^2 values of the respective models, reinforcing the view that these are not good predictor variables.

In this Form three modeling, after the first step one, with the time one oral assessment (RSM1) added to the equation, $R^2 = .39$, F_{inc} (1, 248) = 158.97, $p < .01$**. After step two, with medium of instruction added to the equation, $R^2 = .42$, F_{inc} (1, 247) = 10.60, $p < .01$**. After step three, with a dummy variable for school level added, $R^2 = .42$, F_{inc} (1, 246) = 1.18 (not sig.). After step four, with a dummy variable for student level added (for years one and two), $R^2 = .44$, F_{inc} (1, 245) = 8.87, $p < .01$**. Finally, after step five, with a dummy variable for NET teacher added (for years one and two), $R^2 = .44$, F_{inc} (1, 244) = 0.28 (not sig.). From the sequential F-tests we can conclude that the inclusion of the 'school level' and 'NET teacher' variables in our modeling do not permit us to reject the null hypothesis that for these

two variables the multiple R (and R^2) equal zero. This leads us to conclude that they are not useful predictor variables for form three students.

The final step five model for Form three students with all the predictor variables in the equation , R = .66, F (5, 244) = 38.25, p < .01**.

Form Three Summary Equations

For form three students, we can again formulate a least-squares regression (prediction) equation for the outcome variable *Time two* oral assessment. Equation Z below shows this equation using unstandardised 'Beta' values:

Predicted RSM2 = -.27 + (.47)(RSM1) + (2.43)*(MOI) - (.73)*(D_STU_LV) + (1.77)*(D_STU_LV) + (.27)*(D_NET)* (Z)

This same equation may be expressed in standardized Z-scores, as shown in equation A below:

$$Predicted\ Z_{RSM2} = 0 + (.50)*(Z_{RSM1}) + (.23)*(Z_{MOI}) - (.07)*(Z_{D\_SCHLV}) + (.16)*(Z_{D\_STU\_LV}) + (.03)*(Z_{D\_NET})$$ (A)

4. Form Four

As with the whole sample analysis and Forms one and three, the respective correlations between the outcome variable and the predictor variables were initially calculated, with form four students producing coefficients ranging from r = .69 (RSM1) to r = -.16 (NET). A full correlation matrix for Form four students is shown in Table 41 below.

Table 41. Intercorrelations between outcome variable (RSM2) and predictor variables (Form four, n = 120)

IVs	RSM2 (DV)	RSM1	MOI	d_sch_lv	d_stu_lv
RSM1	.69				
MOI	.29	.38			
d_sch_lv	.29	.38	1.00		
d_stu_lv	.56	.43	.00	.00	
d_net	-.16	-.13	-.18	-.18	.21

Table 41 above shows that the correlations between the outcome variable and the predictor variables for Form four students were somewhat different from those of the From one and Form three students. In this instance, the RSM1 correlation was higher at r = .69 (Form one r = .62, Form three r = .62), while the correlation for the MOI variable was much lower at r = .29 (Form one r = .45, Form three r = .40). In addition, the correlation for 'Net teacher' was much lower and negative at r = -.16, p = .04*. From this initial correlation matrix it would appear that the strongest predictors for Form four students are likely to be 'Time one oral assessment' and 'student level' while the predictor 'NET teacher' is likely to be weak. To explore this further, a summary of the unstandardised Beta coefficients (*B*), standard errors (*SE B*) and standardised coefficients (β) was calculated with the results shown in table 42 below.

Table 42. Summary of hierarchical regression models for outcome variable (RSM2) from predictor variables (Form four, n = 120)

Variable	*B*	*SE B*	β
Model 1			
Time one oral assessment	.61	.06	.69**
Model 2			
Time one oral assessment	.60	.06	.67**
English Medium of Instruction	.26	.63	.03
Model 3			
Time one oral assessment	.45	.07	.50**
English Medium of Instruction	.83	.58	.09
Student level	3.24	.64	.35**
Model 4			
Time one oral assessment	.41	.07	.46**
English Medium of Instruction	.68	.57	.08
Student level	3.73	.65	.40**
NET teacher	-1.49	.55	-.17**

*p < .05 **p = < .01

Table 42 above, shows that in the final model, three of the four predictor variables produced significant standardised Beta coefficients (β). The values for RSM1 and student level were relatively high at .46** and .40** respectively. Of the other two predictor variables (MOI and NET teacher), the β values were low at .08 (not sig.) and -.17** respectively. As with Form three students, the Standard Errors were high relative to the unstandardised Beta coefficients. This suggests that for Form four students, MOI and NET teacher were not good predictor variables for the regression model. To explore this further, the relative statistics for the multiple correlation coefficient (R), R Square (R^2), the adjusted R Square (Ad R^2) and the R^2 were also calculated and are shown in table 43 below.

Table 43. Sequential regression model summary showing relative R values (Form four)

Model	R	R^2	Adj R^2	R^2 Change
1	.69	.47	.47	.47
2	.69	.47	.46	.00
3	.75	.57	.56	.10
4	.77	.59	.58	.03

Table 43 above shows that the percentage of variation as shown by the adjusted R^2 statistic, increased from 47% in the first model (Form one = 39%, Form three = 39%), to 58% in the final model (compared to 48% in Form one students and 44% in Form three students). The inclusion of the variable 'MOI' in model two did not increase the Adjusted R^2 value but the inclusion of 'NET teacher' in model four did increase the Adjusted R^2 value albeit by a small amount. In common with Form one and Form three students, the most significant predictor variable was Time one oral assessment (RSM1) but on the other hand, the medium of instruction (MOI) of Form four students was not a good predictor of the outcome variable. The inclusion of NET teacher into the model did slightly increase the adjusted R^2 statistic although the β value was low and negative (-.17**).

We can conclude that in the form four modeling, after the first step, with the time one oral assessment (RSM1) added to the equation, $R^2 = .47$, F_{inc} (1, 118) = 104.86, $p < .01$**. After step two, with medium of instruction added to the equation, $R^2 = .47$, F_{inc} (1, 117) = 0.68 (not sig.). After step three, with a dummy variable for student level added, $R^2 = .57$, F_{inc} (1, 116) = 25.62 $p < .01$**. Finally, after step four with a dummy variable for NET teacher added (for years one and two), $R^2 = .59$, F_{inc} (1, 115) = 7.47, $p < .01$**.

The final step four model for Form four students with all the predictor variables in the equation , R = .77, F (5, 115) = 41.95, p < .01**.

7. Form Four Summary Equations

From this analysis, a least-squares regression (prediction) equation can again be formulated for the outcome variable *Time two* oral assessment for form four students. Equation B below gives this equation using unstandardised Beta values:

$$Predicted\ RSM2 = .03 + (.41)*(RSM1) + (.68)*(MOI) + (3.73)*(D\_STU\_LV) - (1.50)*(D\_NET) \quad (B)$$

This may also be expressed in standardized Z-scores, as shown in equation C below:

$$Predicted\ Z_{RSM2} = 0 + (.46)*(Z_{RSM1}) + (.08)*(Z_{MOI}) + (.40)*(Z_{D\_STU\_LV}) - (.17)*(Z_{D\_NET}) \quad (C)$$

These models would account for some 58% of the variation, the remainder of which would be the residual or error.

IV. SUMMARY AND CONCLUSIONS

1. NET Effect

The multiple regression analysis conducted on the four sample groups (whole sample, Form one, Form three and Form four) produced somewhat inconclusive results, as can be seen from table 44 below.

Table 44. Summary multiple regression table showing key results of the predictor variable 'NET teacher' (whole sample, F1, F3 and F4)

	n	NET teacher sig. β	multiple R	model total adj. R^2
whole sample	790	.09** (yr1)	.73	.52
		.03 (yr 2)		
F1	420	.11**	.70	.48
F3	250	.03	.66	.44
F4	120	-.17**	.77	.58
mean			.71	.50

** p = < .01

Firstly, the models were able to account for, on average, 50% of the variation, with adjusted R^2 values ranging from .44 to .58. In the case of Form four students, a relatively high 58% of the variation could be explained, whilst in the case of Form three students, only a relatively modest 44% could be attributed to the predictor variables. As regards the significant standardized Beta coefficient (β), the values were in all cases relatively small, although in two cases the values were statistically positively significant (whole sample [year1] β = .09**; Form one β = .11**). On the other hand, there was one negative coefficient, namely Form four (β = -.17**). It is difficult on the basis of these results to make inferences from the sample to the

population. There is some evidence of a marginal positive effect however, but this would need to be substantiated through more research, preferably over a longer period of time and with a larger sample size.

2. Other Predictor Variables

The other key predictive variables used in this modeling were: Time one oral assessment, medium of instruction, school level and student level. Table 45 below gives a summary of the standardized Beta coefficient values (β) for the other key predictor variables for the final sequential model.

Table 45. Summary of coefficient values (β) for other predictive variables for the outcome variable RSM2 (all groups)

Group	Time 1 Oral Assessment	MOI	School Level	Student Level
Whole sample (n = 790)	.34**	.29**	.10**	.41** .26**
F1 (n = 420)	.44**	.48**	-.25**	.19**
F3 (n = 250)	.50**	.23**	-.17	.16**
F4 (n = 120)	.46**	.08**	.40**	-.17**

** $p < .01$

As we can see from the above table, most of the standardized Beta coefficient values are significant at the $p < 0.01$ level with only one case (F3, 'school level') not significant. Furthermore, the values are generally quite high, particularly those of 'Time one oral assessment' and 'MOI' suggesting that these are good predictor

variables. This is in fact not surprising given that these two variables formed an important part of the theoretical basis of this (sequential) regression (see p.135-6). The variable 'school level' was not found to be such a good predictor, but on reflection this might be due to the fact that this variable is not entirely independent as it is influenced by MOI (see footnote five).

The next chapter reports briefly on a multilevel modeling (MLM) analysis that was conducted on the same data. The report in chapter seven will be limited to the main findings of this analysis which will be compared to the multiple regression analysis detailed in chapter six above. There are a number of theoretical issues involved in comparing these two types of analysis but these will only be touched upon briefly.

Footnotes

[1] Field (2000) notes that the terms 'dependent' and 'independent' variables are frequently (and in his view incorrectly) used in regression analysis. Since variables are used simultaneously and without control, it is not strictly speaking correct to label them in this way. Field (op cit) uses the terms *predictors* instead of 'independent variables' and *outcomes* instead of 'dependent variables' (p103).

[2] The adjusted R^2 statistic indicates the loss of predictive power or *shrinkage* (Field, 2000). Ideally the value would be the same or very close to, the value of R^2 (non-adjusted). A slightly smaller adjusted R^2 means that if the model was based on the population (rather than the sample), the percentage difference is the amount of reduced variation explained in the model.

[3] It should be noted that the R^2 statistic of 0.4535 (45.35%) is derived from the corrected Rasch calibrated scores. When the R^2 statistic was calculated based on raw scores, this figure was slightly lower at 0.4403 (44.03%). This calculation was based on the sample taken as a whole, but a similar slight difference (in favour of Rasch calibrated scores) was noted with F1, F3 and F4 students.

[4] See Appendix IV for a full summary of 'whole sample' correlations.

[5] It is the case that there are, for example, *no* band five EMI schools and indeed the overwhelming majority of EMI schools are band one or two. (There *are* of course band one CMI schools, but they are proportionally less in number and the general correlation between EMI schools and high school banding is well known in Hong Kong. This is one of the 'problems' in introducing educational reform in the Territory).

[6] This analysis of assumptions was carried out on all the sample groups, although only the 'whole sample' group is reported in the thesis. It should be emphasised that the other sample groups (F1, F3 and F4) produced similar results.

[7] Tabachnick and Fidell (1996) refer to this as *sequential* multiple regression.

[8] There is also a wealth of literature for example, on the selective secondary schooling system in the UK in which students in 'secondary modern' schools consistently under performed in their subjects whereas their 'grammar school' counterparts fared much better.

[9] Teachers were asked to rate their students in *both* of the two years of this study.

[10] As we have seen, the category 'both' is more complex than might be initially expected. The different possible combinations of this deployment, together with the variations within the 'NET' mode of deployment are discussed more fully in Chapter 4, Section II.2.

[11] In the new, revised school level variable, both 'high' and 'medium' level schools were combined together, while 'low' level schools were not included.

[12] It should be remembered that the variable 'student level' is an arguably subjective teacher assessment of the student's ability i.e. there is no objective data to substantiate this. It is nevertheless considered to be a valid category as teachers' intuitive and day-to-day knowledge of their students' ability tends to correlate highly with external assessment scores.

[13] In years one and two, students were categorised as either high, medium or low ability, making a total of six categories. These six were rationalised into just two: high level - not high level.

CHAPTER 7

MULTI LEVEL MODELLING

Having obtained and employed Rasch calibrated scores (Chapter 5) to analyse the data through constructing ordinary least squares (OLS) models (Chapter 6), this study proceeds to further refine the statistical analysis through the use of multi-level models (MLM). The following section gives a brief rationale for employing this approach which will then be followed by a description and analysis of the results.

I. THE MULTILEVEL MODEL

Paterson (1991) asserts that there are two reasons why multi-level modelling (MLM)[1] offers two distinct advantages over and above ordinary regression that researchers should be aware of. The first of these is 'substantive' in that ordinary regression in this current study does not take schools into account. We should remember for example, that the students whose oral assessment scores are being analysed come from some forty nine different schools. Through the use of MLM forty nine separate equations are estimated and the 'within school' component (i.e. the comparison of students' attainment gain attending *the same* school) is separated from the 'between school' component (thus taking into account differences *between* schools). Kreft and De Leeuw (1998) similarly refer to micro-level measurements (i.e. students) and macro-level contexts (i.e. schools)[2]. Kreft and De Leeuw note the importance of recognising that students within a school are often more alike than, for example, their peers in another school.[3] By acknowledging that there are *hierarchies of data*[4] this nested structure of the data can be taken into account, which if ignored

could have consequences for the validity of the results. In analysing data, if the statistical model developed by the researcher does not acknowledge a hierarchical structure of the data, the conclusions and inferences drawn from the analysis might be misleading. The second advantage of MLM referred to by Paterson (1991) is a technical one. In OLS analysis, since the standard errors on the parameters are 'misleadingly small', this means that the confidence intervals are restricted or conservative because the schools introduce an extra random component, and thus an extra degree of uncertainty. In OLS analysis, we might erroneously attribute to a particular variable, a large difference that might easily have arisen as a result of differences between two different schools. Any analysis that ignores schools would understate the effects of chance and the way to overcome this is through the use of multi-level models. Thus multilevel models appeal to the researcher not only on theoretical grounds, but they also may also provide greater insights into the processes generating the data (Rice and Leyland, 1996). The section below briefly describes the extension of OLS models to multi-level models in which the shortcomings referred to above can be addressed.

1. From OLS to MLM

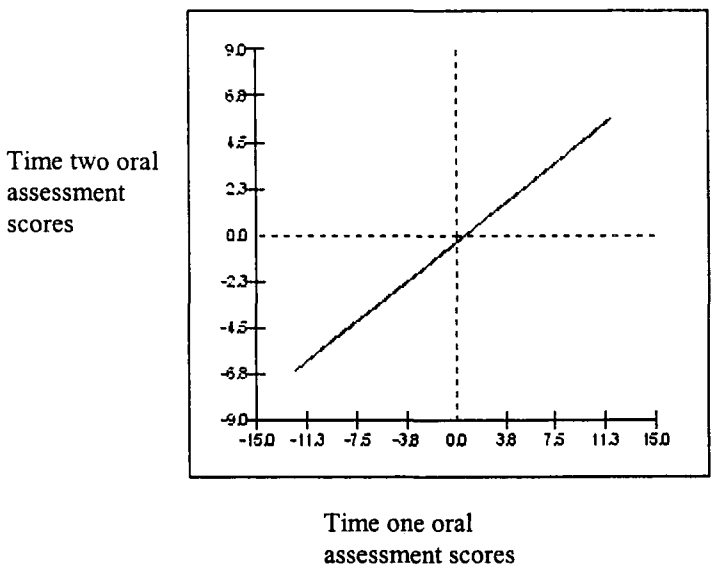
At a basic level, a multi-level model is an extension of an OLS regression model. Ordinary regression would estimate an equation by pooling all the cases together, expressing in this case, the time two oral assessment score as a linear function of the time one score, as follows:

$$y_i = b_0 + b_1 x_i + e_i \qquad (7.1)$$

Here, the subscript i represents the i-th pupil's score, while y is the predicted time two oral assessment score and x is the time one oral assessment score. The

intercept b_0 is where the regression line meets the vertical y axis, b_1 is the slope coefficient and e is the error or residual[5]. We can thus estimate how much, on average, a student's time Two oral assessment score increases (or decreases) for a unit change in the Time one oral assessment variable This model, as we have discussed above however, fails to recognise the effect of schools and as we have seen is graphically represented by one single regression line as shown in Figure 21 below.

Figure 21. Graph showing Time one and Time two regression line (Form one)



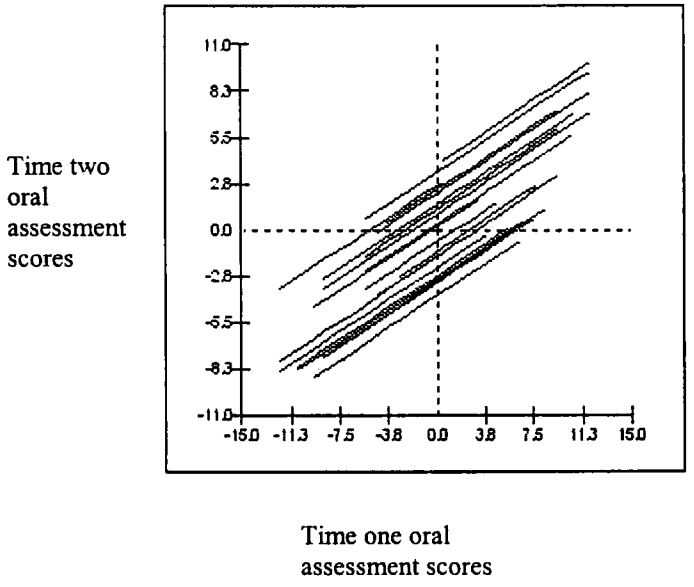
A multi level model however takes into account the school differences so that in equation two below, y_{ij} is the attainment of the ith student in the jth school. The term u is specific to each school denoting that each school has its own contribution.

$$y_{ij} = a + b_1 x_{ij} + u_j + e_{ij} \qquad (7.2)$$

In this equation, u_j and e_{ij} form the *random* part of the model, whose variances, σ_u^2 and σ_e^2 respectively, need to be estimated. The mean intercept *a* and the slope *b* are the fixed parameters of the model.

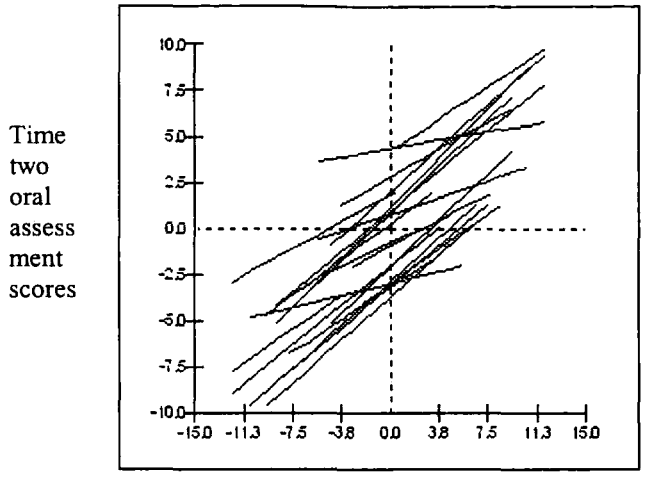
It is useful to build up a graphic picture of the residuals from individual schools and to see how this modelling differs from OLS. By allowing schools to have different intercepts, we can see from Figure 22 below, that each of the school summary lines are parallel and the differences between schools are constant across the range of Time one scores. This is the *simple* level two variation (Rasbash et al, 2000).

Figure 22. Graph showing separate Time one and Time two regression lines for each school (Form one)



Finally, we can build a *complex* level two variation by allowing the school summary slopes to vary as shown in Figure 23 below.

Figure 23. Graph showing Time one and Time two random intercept and slope models



The above example represents a two-level multilevel model of students nested within schools and it is this two-level hierarchy that is adopted for the further analysis of forms one, three and four in this study. We can see from Figure 23 above that the lines of some school slopes are flatter whilst others are steeper. This indicates that the oral English language proficiency gain (Time one to Time two) of students is greater in some schools than it is in others. By adding more predictor variables to the model we can investigate these patterns within schools. However, this analysis is not within the scope of this thesis which will be restricted to a simple comparison between the OLS and MLM models. The results of this analysis are described in the following section.

II. RESULTS

Some analysts (e.g. Tymms, 1997) conclude from their research that results derived from OLS analysis and MLM procedures differ very little except that the errors on the parameters derived from MLMs tend to be larger than those from OLS and it is the MLM errors that should be noted. Multilevel modelling analysis was conducted on Forms one, three and four using the same variables as those used in the OLS analysis. Comparisons were made between the two types of analysis to see if the amount of variance explained be the respective models differed substantially.

1. Form One Analysis

Table 46 below shows a summary of the Form one MLM analysis, comparing the β coefficients at all stages with those derived from OLS analysis and showing the variance for random school and student residuals.

Table 46. Summary MLM table with OLS comparisons (Form one)

	null model	1st	2nd	3rd	4th	Full	OLS
Fixed							
Constant	-0.41	-0.33	-1.56	-1.26	-2.59	-2.78	-2.50
		(0.59)	(0.62)	(0.64)	(0.67)	(0.67)	(0.37)
RSM1		0.48	0.47	0.48	0.37	0.36	0.39
		(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.04)
MOI			3.18	6.72	6.36	6.43	4.70
			(0.97)	(2.18)	(2.18)	(2.14)	(0.95)
School				-3.80	-3.06	-3.23	-2.39
Level[6]				(2.16)	(2.15)	(2.12)	(0.95)
Student					1.90	1.91	1.90
Level[7]					(0.33)	(0.33)	(0.39)
NET						0.44	1.10
						(0.37)	(0.35)
r^2							0.48
Random							
Student	13.53	8.17	8.16	8.08	7.48	7.48	
	(0.96)	(0.59)	(0.59)	(0.76)	(0.54)	(0.54)	
School	12.17	6.56	4.34	4.49	4.25	4.00	
	(4.00)	(2.23)	(1.54)	(1.58)	(1.50)	(1.42)	
Drop (student)		40%	40%	40%	45%	45%	
Drop (school)		46%	64%	63%	65%	67%	

Note: Figures in parenthesis are Standard Errors of the given statistic.

In both OLS and MLM models, the final beta coefficients of the variables RSM1, MOI, and student level were all statistically significant, as evidenced by the fact that the standard errors in brackets are less than half of the betas. The MLM full model produced a beta coefficient of 0.44 (SE 0.37) i.e. not significant, whereas the value of the NET beta in the OLS modelling was significant at 1.10 (SE 0.35). This suggests that in Form one students, the evidence to support the notion of a NET effect is at best somewhat weak. In the second multilevel model above we can see that the inclusion of the variable MOI significantly reduces the school level variance from 6.56 (SE 2.23) down to 4.34 (SE 1.54) indicating that a lot of the variance was

attributable to the school level two random residuals. If we then compare the value of the MOI beta coefficient in the 'full' multilevel model with that of the OLS model we can see that the value of the former is much higher but both are significant.

2. Form Three Analysis

The results of the Form three MLM analysis again produced results similar to the OLS procedure with 'RSM1', 'MOI' and 'student level' all producing significant beta coefficients. As with the OLS analysis, there was no significant NET effect on Form three students. Table 47 below gives full results.

Table 47. Summary MLM table with OLS comparisons (Form three)

	null model	Full	OLS
Fixed			
Constant	1.12 (0.90)	0.90 (0.95)	-0.27
RSM1		0.39 (0.06)	0.47 (0.05)
MOI		2.48 (1.83)	2.43 (0.72)
School Level		-0.25 (1.84)	0.73 (0.74)
Student Level		2.31 (0.51)	1.77 (0.59)
NET		0.17 (0.64)	0.27 (0.52)
r^2			0.44
Random			
Student	15.33 (1.42)	10.54 (0.98)	
School	12.60 (4.73)	5.18 (2.05)	
Drop (student)		31%	
Drop (school)		59%	

Again we can see that there was a large drop of 59% in variance at the random school level two between the null and the full models. In the MLM therefore, a large part of the variance at this level has been explained between the first (null) and the last (full) model. The remaining school level variance of 5.18 (SE 2.05) suggests that

the oral English language proficiency gain of students differs considerably according to the school they attended. As with the Form one students, the medium of instruction explains a significant amount of variance at the fixed level of the MLM, and for a significant percentage of the variance in the OLS model. There was very little difference between the results of the NET variable as modelled by OLS and by MLM and it was not found to be significant in either case.

3. Form Four Analysis

MLM of the Form four students' results did not differ greatly from the OLS analysis, as indicated by the final beta values which were very similar and are shown on Table 48 below. In both OLS and MLM significant amounts of variance were computed for the variables RSM1, student level and NET teacher. The variable MOI was not statistically significant in either the OLS or MLM analysis. As with the Form one and form three analysis, a large amount of the residual was attributed to the random school level two. Although some 64% of the variance was reduced between the null and the final model, with this group of Form four students, most of the variance was in fact at the student level. Another important point to note is that the NET variable is statistically significant, but has a negative value. This suggests that at Form four, the NET variable does not explain any of the variance in the model, but it is in fact the local teacher who does so. Thus for Form four students in this study there is no evidence of any NET effect in oral proficiency gain. It might well be that at this age and stage of their education, the students are more focused on their public examinations (the preparation for which local teachers are arguably best suited), and students are not interested or motivated by 'fun activities'. Further analysis of the data, including residual analysis, would be required to resolve this issue.

Table 48. Summary MLM table with OLS comparisons (Form four)

	null model	Full	OLS
Fixed			
Constant	1.98 (0.66)	0.11 (0.72)	0.03 (0.62)
RSM1		0.44 (0.07)	0.41 (0.07)
MOI		0.64 (0.89)	0.68 (0.57)
Student Level		3.72 (0.63)	3.73 (0.65)
NET		-1.52 (0.60)	-1.49 (0.55)
r^2			0.59
Random			
Student	16.46 (2.19)	6.87 (0.91)	
School	2.00 (1.64)	0.72 (0.62)	
Drop (student)		58%	
Drop (school)		64%	

Having completed the statistical analysis, the next chapter will move on to discuss the constraints and possible limitations of this research. Where appropriate, and in light of the acknowledged limitations, recommendations will be made that might help future researchers avoid some of these pitfalls in the design, application and analysis of similar projects.

Footnotes

[1] Paterson (1989) also notes that researchers use the term 'multi-level regression' as well as 'multi-level modelling'. In this thesis, the two are considered to be synonymous and the latter will be used throughout.

[2] Hence the term 'contextual models', used by some analysts.

[3] Rasbash et al (2000) note that in medical studies, centres often differ in terms of patient care, case mix, etc.

[4] 'Hierarchical linear modelling', is also a term used synonymously with MLM.

[5] The residual is assumed to be normally distributed with a mean of zero and a constant variance.

[6] The variable 'school level' is not to be confused with the level two school residuals under analysis in MLM. This variable, it will be remembered, is a dummy and is related to the banding of the schools in the sample (high, medium and low).

[7] As with the above, the 'student level' variable should not be confused with the level one residuals in MLM. It is the dummy variable representing the teacher's view of whether the student is of high, medium or low ability.

CHAPTER 8

LIMITATIONS AND RECOMMENDATIONS

This chapter considers some of the limitations to the present study. Inevitably with a study of this nature that involves the active participation of many external agencies, in hindsight there are a number of improvements that could ideally have been made to this study. This study, it will be remembered, was part of a larger scale research project that involved the cooperation and collaboration of several researchers and research assistants. In addition, this investigation into the oral proficiency of Hong Kong secondary students also involved the cooperation of some forty nine schools, two or sometimes three assessors from those schools and of course those students being assessed. Finally, the views, agendas and constraints of several external bodies such as ED, SCOLAR and the HKIEd have had to be considered in the planning and execution of this project. In so doing, compromises have to be reached leading to decisions that are less than perfect. Certain limitations of this study can be attributed to such constraints, whilst others can be attributed to decisions and techniques undertaken by the researcher that have raised questions and problems. All of these are considered and discussed in the following section.

I. LIMITATIONS OF PRESENT STUDY

1. Lost and Missing Data

One of the key problems encountered during the course of this project, which to some extent may have compromised the findings concerns lost data. In the first

administration of the oral assessment, some 1,426 were interviewed while in the second administration the number totalled 928. Despite this relative large sample number, there was a lot of lost data so that in the multilevel modelling stage for example, there were only 790 cases with full sets of data. Whilst much of the analysis in this study was conducted using the statistical package SPSS this was not a problem since SPSS allows the analyst to exclude cases either pairwise or listwise. However, the MLWin software package used for the multilevel modelling only permits the entry of *full* data sets, hence the severely reduced number of cases in this analysis. The post test assessment was administered to five hundred students less than the pre test, which apart from frustrating the researcher also means that this very wasteful on resources.

In addition to the above wastage (some of which is always inevitable in this type of research), many potentially useful variables were not completed due to schools not submitting the necessary information. The variable 'gender' for example is one that would have been useful to include in the analysis, especially in light of the recent world-wide debate on the widening gap in attainment between girls and boys. Other instances of variables that were finally excluded from the analysis were 'number of hours taught by the NET' plus a range of other assessment results such as HKAT scores and listening test results[1]. It is also a truism that the smaller the sample size the bigger the threat to reliability. Similar research in the future might be able to pre-empt this loss of data through the implementation of the following recommendations:

Recommendations:

- The inclusion of a member of ED in research teams might not only improve access to schools[2], but could also bring pressure to bear on school authorities to supply the necessary data.
- The inclusion of one or more NET teacher and local teacher would help the research team gain valuable insights.
- The careful design of forms and administrative procedures would help ensure that information such as 'gender' does not get accidentally overlooked.
- Trying to ensure that *all* data required from schools and teachers is done in one go rather than in several rounds might also encourage the supply of more data and improve overall efficiency.

2. Time Scale

Since this project was conducted over a two-year period, the time imposed severe constraints. Once such a project is under way and schools, teachers and students are involved in training workshops and the administration of assessments, the project develops a momentum of its own that can not be controlled (e.g. school holidays, availability of teachers/students, public examinations, etc.).

Recommendations:

There is therefore a need to ensure:

- a longer lead-in time for planning, and;
- a longer overall research time line[3].

A further point on time scales concerns the measurement of language proficiency gain. While there is no doubt that proficiency gain does occur over a two-year period, as this study has shown, such gain is notoriously difficult to measure over short periods. Further, to try to attribute language gain to variables such as a NET teacher and measure any 'NET effect' is more realistic over a longer period of time. These issues could be addressed by the above recommendations.

3. Sample Size

One limitation of this study is that the sample size involved. It is acknowledged that due to the wastage of data described above, one must be cautious in making sweeping inferences from this small sample size. A further note of caution should also be sounded regarding the generalisability of these results to the population as a whole.

4. Follow-up Study

The issue of restricted time scales has already been alluded to above. In order to build on this existing body of research it is therefore suggested that a small-scale follow-up study be conducted in order to complete the analysis of missing data such as gender. At the same time, additional data on the longer term effects of the students' involvement (or not) with NET teachers could be obtained. This might reveal for example, whether there is there any relationship between students being taught by NET teachers and their future public examination results, particularly in spoken English.

5. Inter-rater Reliability

It has been acknowledged that a possible limitation of this study lies in the issue of inter-rater reliability. Due to time and resource constraints, it was not possible to conduct a concurrent inter-rater reliability study during the course of the project as one would ideally wish. Instead, a small-scale post hoc inter-rater reliability was carried out which to some extent has allayed concerns in this area, but they do nevertheless exist. It is recommended such a process be built into the design and methodology of any such similar future project to ensure greater confidence in the reliability of the results.

Other measures that could be taken to increase confidence in the reliability of this type of study include:

- ensuring that there are *two* assessors for every interview: one to act as 'assessor' and the other as 'interlocutor'[4], although this has obvious resource and cost implications;
- building into the research methodology a means whereby the oral assessment interviews could be better monitored either by having a small team of monitors or sit in on assessment interviews[5] and/or to randomly sample the recorded audio tapes.

6. Technical Issues

This thesis was the result of a purely quantitative data analysis which it may be argued has limitations. It is generally considered best to triangulate a quantitative

analysis with qualitative data to help ensure a more balanced and fuller picture of the issues under investigation. Added to this there are technical questions regarding the use of the Bell Curve itself (Carroll, 1997; Glymour, 1997). Is it possible therefore to make general inferences on a population based solely on the quantitative analysis of a small sample? In response to this possible limitation, firstly it is argued that the researcher has drawn on a variety of quantitative research sources, particularly in the literature review and methodology design. Secondly, as regards the more technical issues, it is felt that the oral assessment instrument developed for this study was suitably robust and that the data analysis techniques were rigorous in their scope and detail. Given the limitations on the data base (sample size, restricted number of predictor variables), it is felt that further quantitative analysis is unlikely to shed more light on the question of a NET effect on proficiency gain.

7. Comparative Literature

One noticeable limitation of the current study is the lack of related comparative literature with which the results could be compared. It has not been possible to date to find a body of literature concerning studies in similar contexts that have evaluated the impact of native-speaking English teachers on the language proficiency gain of students. To the knowledge of the author, there are projects in other countries in which NETs are involved in teaching English in secondary schools: the Japan Exchange Teaching (JET) scheme in Japan is one such example. Whilst there is a body of literature on the JET scheme, searches have failed to reveal any studies on programme evaluation or proficiency gain. Much of the literature is directed towards affective factors such as motivation, culture conflict (e.g. Barratt and Kontra, 2000), a whole host of issues surrounding methodology and practical guides

(e.g. JALT Journal), and issues such as native English-speakers living in a foreign country (e.g. Scully, 2001; Kramsch, 1993).

Recommendations:

A number of governments around the world are investing considerable resources into recruiting and employing native-speaking English teachers in their schools with a view to trying to raise language standards. So too are other non-governmental organisations such as Voluntary Services Overseas (VSO). There is therefore a need to conduct a study into the evaluation of any such programmes with reference to language proficiency gain. Whilst there is no doubt that these programmes have many positive outcomes, particularly in terms of student motivation (e.g. Luk, 2001), multi-culturalism and teacher education, not enough is yet known on how the programmes impact upon the students' language development. A comprehensive study on this issue which would seek the co-operation of the different parties involved (for example, the education departments of interested governments, teachers' representatives and students) is urgently needed.

II. GENERAL RECOMMENDATIONS

1. Monitoring Language Standards Over Time

As previously mentioned, the need for this study and indeed the MENETS project arose not only to monitor the effectiveness of NETs but also implicitly because of concern over perceived falling English language standards. It will be remembered that the claim being made is that the deployment of the NET Scheme will address this latter problem. Yet as we have seen, there is conflicting evidence on

the issue of falling standards, with neither side of the debate having to date built a solid case to substantiate their claims.

It is recommended therefore that measures be put in place to monitor the English language proficiency of students in Hong Kong secondary (and primary) schools, and it is argued that any such scheme must be a long term, on-going process. To this end, ED and/or its agencies could put out offers to tender for an effective English language monitoring project. This might be done in a number of ways, such as:

- the monitoring of HKEA results over a period of time;
- the development and administration of a battery of language tests by researchers and test developers;
- the funding of a meta-analysis project which would include a range of data, including possibly that used in this current study.

2. Deployment of NETs

During the course of this project, the researcher had many opportunities to visit schools and interact with students NETs and local teachers. There is much evidence that in some schools the NET teacher is successful, is able to motivate the students and has a positive impact on his/her local counter-part and other colleagues. This is not universally true however, which prompts the question as to why many are successful in their results whilst some are not. It is also true to say that the NET teachers are deployed in a variety of different ways with some working in close co-operation with their local counterparts and some working in a more-or-less detached

environment (Luk, 2001; MENETS, 2001; Walker 2001). As the NET scheme involves considerable personal, professional and financial investment on the part of many stake holders, I believe that it is vital to investigate and research this area more fully.

It is recommended therefore that further research by conducted in order to determine: the most effective deployment of NETs; how this valuable resource can best be utilised, and; the most effective means of maximising the teaching learning environment and thus enhancing student language proficiency gain. There is evidence for example that enhancing collaboration between NETS and local teachers has positive effects and that a better shared understanding (by schools, NETs, local teachers and students) of common goals increases efficiency and motivation. This in turn can only have a positive long-term effect on language gain although this is as yet unproven.

Having now fully discussed what are considered by the author to be the practical and technical limitations of this study, let us now turn to the overall conclusions that may be drawn from this thesis and discuss some of the implications. These areas will be presented in the next and final chapter.

Footnotes

[1] For a full range and analysis of other variables included in the MENETS project, see MENETS (2001).

[2] Although in this study the overwhelming majority of schools were very co-operative and helpful in acceding to the researchers' requests.

[3] There is currently another NET evaluation project under tender and it is understood that the principal investigators are planning on a *three year* time line for a *two year* longitudinal evaluation project. This would seem to be the correct strategy.

[4] This is already common practice in many (high stakes) public examinations including those administered by the HKEA and UCLES.

[5] This proposal however might pose other problems in that assessors might feel threatened and/or their close relationship with their students – a distinct advantage of the existing set-up – might be undermined.

CHAPTER 9

SUMMARY FINDINGS AND CONCLUSIONS

I. CONCLUSIONS AND DISCUSSION

1. General Comments

Lee et al (1998) note that trying to measure progress in a second or foreign language is 'one of the most persistent yet unanswered (some would say unanswerable) questions. Yet not attempting to tell how much gain a person is able to make after taking a course of study is indefensible' (p.2). Yet, as Lee et al (ibid) point out, it is necessary to continue to strive to do so in order to justify the continued funding of particular programmes. This study would certainly concur with these views.

This thesis has given an account of the investigation into the English language oral proficiency gain of secondary school students in Hong Kong over a two year period from the beginning of the 1998-1999 academic year, up to the end of the following academic year (1999-2000). In particular, the thesis has been concerned with examining whether any such language gain may be attributable to students having been taught by NETs – the so-called 'NET effect'. This thesis has also considered what other variables could be considered strong predictors of the outcome variable oral English post test assessment. It is felt that the analysis presented in this thesis has been rigorous and in its attempt to measure oral English language proficiency gain in relation to NETs in Hong Kong.

As we have seen from the literature, within Hong Kong there is a vocal body of opinion that questions the effectiveness and rationale underlying the deployment of NETs in Hong Kong schools. Interestingly however, the voices of discontent come mainly from other (local) teachers and indeed from some academics, but as Luk (2001) notes, the negative reaction does not by and large come from the students themselves[1]. In fact, Luk (ibid) concluded from her study that "...the majority of the respondents were in favour of being taught by the NETs ... the contact with the NETs was something they valued and something they had positive expectations of" (p31). Students certainly value the native-speaker largely because of what they bring with them naturally – their authenticity (Barratt and Kontra, 2000), and the fact that students are *forced* to speak in the target language because the NET usually does not speak the students' L1. The positive impact of NETs on their students is certainly well documented, but the results of many qualitative research projects are not able to establish a direct link between students' positive viewpoints (which arise mainly from the administration of questionnaires, but also include ethnographic research involving classroom observations) and measured language proficiency gain. This lack of quantitative data to support the view that the deployment of NETs in the English language classroom necessarily leads to language proficiency gain *in addition to* that which would normally be predicted if students were taught only by local teachers brings us back to the 'common sense' notion referred to in chapter one (p.3)[2].

As a predictor variable, this study has not been able to establish a strong link between the deployment of a NET and oral English language proficiency gain. It must be emphasised however that the issue of measuring oral English language proficiency gain is widely considered to be problematic, particularly if the pre and

post measurements are conducted within a relatively short period of time (in the case of this study considerably less than two years[3]). To the knowledge of the author, this has not to date been successfully achieved and indeed there is much anecdotal evidence to suggest that a longer period of time would be required to measure any proficiency gain in speaking English *and* to establish strong predictor variables[4].

As a result of detailed analysis (chapter 4), it was felt that the data gathered from the oral assessments produced distributions that were more or less normal, with high correlations between the individual assessment components (.82 - .89 in the pre test; .82 - .88 in the post test). Technical aspects such as skewness and kurtosis were found to be within acceptable limits although the distribution of the Form four students' scores was marginal (see Table 15, p. 91). The internal consistency was also found to be high (.97 in both pre and post tests) and this together with the post hoc inter rater reliability study conducted on a random sample of interviewees (see Pp. 68-69), suggest that the instrument used in this research was reliable. In the light of this preliminary analysis, it was felt that the statistical analysis employed and described in this thesis was appropriate for the given data. The theoretical construct on which the instrument was developed together with informal and formal[5] feedback from assessors and students also indicate that it had good validity. An analysis of the distribution of scores by school level, district and medium of instruction tend to support the methodology of the sample group which was a randomised, stratified sample as had been designed to ensure a representative of the population (see chapter 4). It is the view of the author from this analysis that the oral assessment instrument developed for this project was able to measure oral English language proficiency with a good degree of reliability and validity.

2. Measuring Proficiency Gain

Paired samples t-tests revealed that with the whole sample group, the Form three group and the Form four group there were significant increases in the assessment scores between the two times, but there was no significant increase for Form one students. As suggested previously, this could well be due to the fact that in Hong Kong there is a marked drop in students' attainment in all subjects in their first year at secondary school. This is generally true in all schools, but in EMI schools this must be particularly true where students have to adjust to a completely new teaching and learning environment. The gain between Time one and Time two scores is summarised in the table below:

Table 49. Summary t-test results

group	t	sig.	ES
whole sample	t (915) = -7.04	p < 0.01**	0.05
F1	t (915) = -0.68	p = 0.493	0.0005
F3	t (305) = -7.69	p < 0.01**	0.16
F4	t (119) = -6.41	p < 0.01**	0.26

As regards the type of modelling used in this analysis, the OLS analysis produced generally medium to high R^2 values (whole sample = .53; F1 = .48; F3 = .44; F4 = .59) indicating that these models were able to account for roughly half of the variance between the Time one and Time two scores. Whilst the MLM analysis was not able to account for more of the overall variance, the advantage of this technique was that we were able to observe that the random, school level variance remained high even after the full models had been constructed [F1 7.48 (0.54); F3 10.54 (0.98); F4 6.87 (0.91)]. This suggests that most of the variance in the models was in fact

between schools themselves. Further analysis of the residuals would be required to determine in which schools students' gain in oral English language proficiency was greatest. Further MLM analysis might for example also look in more detail at school level and MOI.

3. The NET Effect

There were a number of predictor variables under investigation in this study, first and foremost being the NETs. What light then has been thrown on this question as a result of this study? Is there any evidence from this study to support the idea of a NET effect? In table 44 (p. 165) we can see that the evidence from the OLS analysis for a NET effect is not strong. Whilst there are some significant beta coefficients in the whole sample and Form one groups, the values were not high. In addition in Form three the NET variable did not account for a significant proportion of the variance and in the Form four students the NET variable was in fact negative.

The evidence related to the impact of the NET effect on raising English oral language proficiency suggests that this variable is not a strong predictor of the post test outcome. In the ANOVA analysis, the picture was somewhat mixed although distinct patterns did emerge. In the Time two assessment for example, the average score of students taught by 'both' was consistently lower than the average score of students taught by NETs and by local teachers. This pattern was consistent whether the analysis was conducted by Whole sample, F1, F3 or F4 suggesting that the 'both' deployment is less effective than the other two. This could be the result of teaching consistency, with students being possibly unsettled by not having only one teacher for their English lessons. In the Time one assessment, the picture was less clear, although

again the mean score of students taught by 'both' was generally lower than those taught by just one teacher (whether that was NET or local). In the Whole Sample analysis however, the average score of students taught by NETs was higher than the scores of those by local teachers or by those taught by 'both' although the effect size was very small (ES = .009). The OLS and the MLM analyses did not contribute much towards clarifying the picture of a NET effect. In the OLS analysis, while the beta value for the dummy variable NET for the Whole Sample group was significant, it was also low (ES = .003). For Form one students it was also significant but low (ES = .02), while for Form three students it was not significant. By contrast, with Form four students it was significant but negative (ES = .06). Since the effect sizes were very small this indicates that where there was a NET effect it can only be considered as marginal.

All of this of course is not to say that there *is* no NET effect, merely that over this time period and using these instruments it has not been possible to measure it. It is felt that there is sufficient evidence to suggest that NETs have a positive influence on the motivation to learn and the attitudes towards the English language. Given this, it is highly probable that over a longer period of time and with continued refinement of the instrument used in this study, together with more intensive assessor training such a gain can be measured. Given the subjective nature of oral assessments however, this is always going to be a difficult challenge and it would need the long-term commitment and support of ED to monitor not only oral proficiency gain, but also that of other language areas and skills.

4. Medium of Instruction

The post test analysis of means (see Pp. 99-100) indicates significant differences in the average scores of those students taught in EMI schools and those taught in CMI schools. This was consistently found to be the case whether the analysis was carried out as a whole sample or by individual form. For example, even the pre test results showed significant differences in Form one students' scores between Time one and Time two, as follows: $F(1, 673) = 202.22$, $p<0.01$**. In this example as with other groups, the effect size was large (0.23) suggesting that the medium of instruction has strong influence on students' outcome scores at even the early stages of secondary schooling. In the multiple regression analysis, this was also found to be the case in which the beta values were significant for the Whole Sample group ($\beta = .29$, $p=<0.01$), Form one ($\beta = .48$, $p=<0.01$) and Form four ($\beta = .23$, $p=<0.01$), although at Form four it was not significant ($\beta = .08$, $p=0.24$, not sig.). Again, results from the MLM analysis were similar, except that for Form three, whereas in the OLS analysis MOI was significant [2.43 (0.72)], in the MLM analysis it was not [2.48 (1.83)]. It could well be therefore that due to the 'misleadingly small' standard errors that are noted in OLS (see p. 156 and Paterson, 1991) the difference that we might attribute to MOI is in fact one that has resulted as a result of differences between schools. This is accounted for in MLM.

5. School Level

There were significant differences in both pre test and post test scores between students taught in high, medium and low level schools. As one might predict, the scores of students in higher level schools were on average higher than those of their peers in medium and low level schools. In the Whole Sample group for example, the

F statistic was as follows: F(2, 1421), = 261.01, p=<0.01**. Similar results were observed when the data was analysed by individual form. This was true for both the pre test and the post test results. When the school level variable was explored in more detail in the regression analysis, it was not found to be such a strong predictor variable. This might be due to the high correlation between this variable and EMI (F1 r = .93; F3 r = .69; F4 r = 1.00). Alternatively, results of the MLM analysis suggests that it is not the school level that explains a significant part of the variance, rather it is the school *per se*.

6. Rasch Scale Modelling

The Rasch scale modelling conducted during this research (see Chapter five) was not designed to shed light on any possible NET effect in oral language proficiency. Its purpose, as mentioned previously (Pp. 110-111) was to establish a scale of equal intervals in which item measures, student measures and step difficulty were disentangled. The process then was a means to an end in that the resulting measures were used in the subsequent (OLS and MLM) modelling to investigate strong predictor variables. To this end, it is felt that the RSM succeeded in a number of respects.

The disentangling of the three parameters, item difficulty (δ_i), item steps (τ_j) and person measures (β_n), was achieved and a ruler was successfully created. However, there were altogether only five items involved in the calibration and many of these displayed poor fit in Stage one, and were found to be unstable between Time one and Time two. Consequently, there might have been an insufficient number of stable items for the initial anchoring and adjustment techniques to create a ruler with

better item intervals and steps. Although suitable intervals between student measures seemed to be established, this arguably was not the case with the item measures (Pp. 122-124). The high correlations between raw scores and Rasch calibrated measures was very high (.99) and tended to support the findings of the preliminary raw score data analysis in that assessors were marking holistically rather than using the discrete criteria to discriminate between students. The Rasch calibration tended to support the findings of Wolfe and Chiu (1999a, 1999b) in that the adjustment technique used by Wolfe and Chiu (ibid) and followed in this research does offer considerable advantages. By carrying out this technique, the goodness of fit in both the corrected person measures and the corrected item measures is considerably increased (e.g. uncorrected student measures with fit $\leq \pm$ 2 = 153 (10.62%), corrected student measures with fit $< \pm$ 2 = 134 (9.3%)), and for this reason alone this procedure is to be recommended. The correction technique also improved the fit of four out of the five items which were previously outside the ninety-five percent confidence intervals but were subsequently all within them (see figures 12 on page 118 and figure 13 on page 119). The RSM did not ultimately affect the relative order of item difficulty. In both raw scores and RSM scores, grammar was consistently the most difficult, possibly as a result of teacher intolerance as has already been suggested (p 116). Finally however, whilst the results of this study support the view that theoretically and technically RSM has advantages over the use of raw scores, in this research these advantages were not great when the measures were used for subsequent (OLS and MLM) modelling.

The issue of NETs and English language proficiency gain is one that is likely to continue to be raised in Hong Kong. To measure any gain, particularly in the

spoken language remains a tough challenge for educationists and researchers but through continued attention to methodology, research design and analytical techniques a picture is beginning to emerge. Even if and when there are two NETs in every school in Hong Kong, given the amount of exposure that each child will have to a NET it is a big challenge to all concerned to try to measure the difference they make. It is a challenge that researchers will have to continue to rise to, yet we should do so with the confidence that NETs do make a difference to attitudes and motivation and this must surely over a period of time translate itself into language gain.

Footnotes

[1] From the literature, there seem to be a number of agendas involved in this issue, many of them arguably more to do with politics than genuinely interested in raising English language proficiency in Hong Kong.

[2] There is some evidence of populists appealing directly to public opinion over the heads of the interested parties (in this instance educationists), and using the notion of 'commonsense principles' to do so. Writers of the New right in America (e.g. Chubb and Moe, 1990, 1992) as well as politicians in the UK have done so in recent years (e.g. John Major in his forward to the 1992 White paper on educational reform uses the expression 'commonsense principles').

[3] By the time the measurement instruments were fully developed and the oral assessors were trained, the pre test measurement (scheduled for September/October 1998) did not in fact take place until well into November. There was thus arguably some 'contamination' in the control groups since the students had already been exposed to a Net teacher for a short time before the first measurement was conducted. In addition, the post test measurement (scheduled for June 2000) had to be conducted in many cases in March or April (2000) due to public examinations, students leaving and other constrictions of teachers and schools beyond the control of the MENETS research group.

[4] An example of the difficulty surrounding this issue is well illustrated by the following anecdote. In 1995, the Hong Kong Education Department organised a seminar on overseas English language immersion which was attended by representatives from the HKIEd and overseas universities. One of the speakers, an Australian academic in charge of immersion programmes, reported on the average language gain of students studying in Australia. He informed the seminar that students living in a total immersion context (living with host families) and studying on intensive English language programmes of some 200 hours, improved on their IELTS score, on average, by just 0.5 points!

[5] A formal evaluation was conducted on the oral assessment procedure, consisting of a questionnaire in which Likert scale questions and open-ended responses were included. Results of this formal evaluation from assessors and students were generally very positive and suggest that this instrument was valid. Unfortunately, it has not been possible to include the results of this evaluation in this thesis.

References

Amercan Psychological Association (APA) (2001), *Publication Manual of the American Psychological Association.* (fifth edition). Washington DC: APA.

Árva, V. and Medgyes, P. (2000). Native and non-native teachers in the classroom. *System.* 28 (2000) 355-372.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bacon-Shone, J. and Bolton, K. (1998). Charting multilingualism: language censuses and language surveys in Hong Kong. In M.C. Pennington (Ed.), *Language in Hong Kong at Century's End,* Hong Kong: Hong Kong University Press, pp. 265-277.

Berry, W.D. (1993). *Understanding Regression Assumptions.* Sage university paper series on quantitative applications in the social sciences, 07-092. Newbury Park, CA: Sage.

Bloomfield, L. (1933). *Language.* New York: Holt Rinehart, & Winston.

Bond, T.G. and Fox, C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human sciences.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Boyle, J. (1997). Native-speaker Teachers of English in Hong Kong. *Language and Education,* 11, (3), 163-181.

Braine, G. (Ed.) (1999). *Non-native Educators in English Language Teaching.* Mahwah, N.J.: Lawrence Erlbaum Associates.

British Council. (1988). *Report on the Interim Evaluation of the First Year of the Expatriate English Language Teachers Pilot Scheme.* Hong Kong: The British Council.

British Council (1989). *Final Evaluation Report: Expatriate English Language Teachers Pilot Scheme.* Hong Kong: The British Council.

Burney, E. (1935). *Report on Education in Hong Kong.* London: Crown Agents for the colonies.

Canagarajah, A.S. (1999). Interrogating the 'native speaker fallacy': non-linguistic roots, non-pedagogical results'. In G. Braine (Ed.) *Non-native Educators in English Language Teaching.* Mahwah, N.J.: Lawrence Erlbaum Associates, pp. 77-92.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller Jnr (Ed.) *Issues in Language Testing Research.* Rowley, Mass.: Newbury House (Pp 333-342).

Canale, M. and Swain, M. (1980a). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1 – 47.

Canale, M. and Swain, M. (1980b). *Approaches to Communicative Competence.* Singapore: SEAMEO Regional Language Centre.

Canale, M. and Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer (Ed) *The Construct Validation of Tests of Communicative Competence.* Washington D.C.: TESOL.

Carroll, J.B. (1997). Theoretical and technical issues in identifying a factor of general intelligence. In B. Devlin, S.E. Fienberg, D.P. Resnick and K. Roeder (Eds.), *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve.* New York: Springer-Verlag.

Centre for Applied Linguistics (1989). *Rating Scale for CAL Oral Proficiency Exam (COPE) and Student Oral proficiency Assessment (SOPA).* Washington, D.C.: Centre for Applied Linguistics.

Chan, A., Hoare, P. and Johnson, K. (1997). *English Medium Instruction in Secondary One and Two in Hong Kong Schools: An Evaluation of Policy and Implementation.* Report of a Hong Kong Bank Language Development Fund Project. Hong Kong: Language Development Fund.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA.: MIT Press.

Clark, J. L. (1975). Theoretical and technical considerations in oral proficiency testing. In R. Jones and B. Spolsky, *Testing Language Proficiency,* (pp. 10-24). Arlington, Virginia: Centre for Applied Linguistics.

Cohen, J. (1988). *Statistical Power Analysis for the behavioral Sciences.* Hillsdale, NJ: Erlbaum.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly,* 33, (2) 185-209.

Crystal, D. (1985). *A Dictionary of Linguistics and Phonetics.* Oxford: Basil Blackwell.

Davies, A. (1991). *The Native Speaker in Applied Linguistics.* Edinburgh: Edinburgh University Press.

Davies, A. (1995). Proficiency of the native speaker: what are we trying to achieve in ELT. In G. Cook and G. Seidlhofer (Eds.) *Principle and Practice in Applied Linguistics.* (pp.145-157). Oxford: Oxford University Press.

Dey, I. (1993). *Qualitative Data Analysis.* London and New York: Routledge.

Edge, J. (1988). Natives, speakers and models. *JALT Journal,* 9, (2) 153-157.

Education Commission. (1984). *Education Commission Report No. 1.* Hong Kong: Hong Kong Government Printer.

Education Commission. (1986). *Education Commission Report No. 2.* Hong Kong: Hong Kong Government Printer.

Education Commission. (1990). *Education Commission Report No. 4.The Curriculum and Behavioural Problems in Schools.* Hong Kong: Hong Kong Government Printer.

Education Commission. (1996). *Education Commission Report No. 6.* Hong Kong: Hong Kong Government Printer.

Education Department. (1997). *Guidelines on the Duties of NETs Appointed Under the Enhanced NET Scheme.* Hong Kong SAR: Education Department.

Education and Manpower Bureau. (2002). *Education Statistics.* Hong Kong: Hong Kong Government Printer.

Educational Research Establishment (1991). *Final Report on the Management of the Expatriate Teachers in the Expatriate Language Teachers Modified Scheme.* Hong Kong: Government Printer.

Ellis, R. (1986). *Understanding Second Language Acquisition.* Oxford: Oxford University Press.

Field, A. (2000). *Discovering Statistics Using SPSS for Windows.* London: Sage Publications.

Glymour, C. (1997). Social statistics and genuine inquiry: Reflections on the Bell curve. In B. Devlin, S.E. Fienberg, D.P. Resnick and K. Roeder (Eds.), *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve.* New York: Springer-Verlag.

Hirvela, A. and Law, E. (1991). A survey of local English teachers' attitudes towards English and ELT. *Institute of Language in Education Journal*, 8, pp 25-38. Hong Kong: Institute of Language in Education.

Holliday, A. (1994). *Appropriate Methodology and Social Context.* Cambridge: Cambridge University Press.

Hong Kong Government. (2001). Promoting Bi-literacy and Tri-lingualism. (On Hong Kong government web Site dedicated to policy addresses). http://www.policyaddress.gov.hk/pa01/eindex.html.

Hopkins, K.D., Hopkins, B.R. and Glass, G.V. (1996). *Basic Statistics for the Behavioral Sciences.* Boston and London: Allyn and Bacon.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J.B. Pride and J. Holmes (Eds.) *Sociolinguistics: Selected Readings.* Harmondsworth, UK: Penguin, pp. 269-293.

Jenkins, J. (2000). *The Phonology of English as an International Language.* Oxford: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an International Language. *Applied Linguistics* 23: 83-103.

Johnson, K. (1985). Language in Education. In Cheng, K.M. (Ed.). Collected Papers on Education in Hong Kong: Submitted to the Education Commission in Response to its First Report. Hong Kong: Hong Kong University Press.

Johnson, K. and Tang, G. (1993). Engineering a shift to English in Hong Kong schools. In T. Boswook, R. Hoffman and P. Tung (Eds.), *Perspectives on English for Professional Communication.* (pp. 203-216). Hong Kong: City Polytechnic of Hong Kong.

Johnson, K. (1995). Critical review of literature on language in education in Hong Kong. In *Education Commission Report No. 6: Part 2.* Hong Kong: Government Printer.

Kachru, B.B. (1985). Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk and H.G. Widdowson (Eds.) *English in the World – Teaching and Learning the Language and Literature.* Cambridge: Cambridge University Press/The British Council, pp. 11-30.

Kachru, B.B. (1994). Monolingual bias in SLA research. *TESOL Quarterly,* 28, 795-800.

Krashen, S.D. (1981). *Second Language Acquisition and Second Language Learning.* Oxford and New York: Pergamon Press.

Kreft, I. and De Leeuw, J. (1998). *Introducing Multilevel Modeling.* London and Thousand Oaks, CA: Sage.

Lai, M. L. (1999). JET and NET: A comparison of native-speaking Enflish teacher schemes in Japan and Hong Kong. *Language, Culture and Curriculum, 12*(3), 191-195.

Lai, K.C. (2002). 'Time ripe for teachers to be fully trained, says Institute'. *South China Morning Post.* October 31, 2002.

Law, E. (1987). 'More cost effective ways to improve English teaching'. *South China Morning Post.* February 3, 1987.

Lee, T., Wylie, E., and Ingram, D.E. (1998). Proficiency gain estimation. Handout prepared for 'Mapping rates of progress in proficiency'. A paper prepared for LTRC Conference.

Linacre, J.M. (2001). *Winsteps Ministeps: Rasch Model Computer Programs.* Chicago, IL: www.winsteps.com.

LLewellyn, L. (Chairman). (1982). *A Perspective on Education in Hong Kong: Report by a Visiting Panel.* Hong Kong: Hong Kong Government Printer.

Lo, T. (1992). 'A socio-cultural framework for a critical analysis of English as a Foreign Language in Hong Kong'. A paper presented to the Annual International Meeting of the ILE. Hong Kong: ILE.

Lo, W.K. (1998). A study on the impact of the Enhanced Native-speaking English Teacher (NET) Scheme on its participating teachers (NETs) in their first year of teaching in Hong Kong. (Un-published M.Ed. dissertation), Hong Kong: Hong Kong University.

Luk, J. (2001). Exploring the sociocultural implications of the native English-speaker teacher scheme in Hong Kong through the eyes of the students. *Asia Pacific Journal of language in Education,* 4 (2), 19-49.

Lung, J. (1999, June/July). A local teacher views the native teacher scheme in Hong Kong. *TESOL Matters,* p.8.

Medgyes, P. (1992). Native or non-native: Who's worth more? *English Language Teaching Journal* 46, (4) 340-349.

Medgyes, P. (1994). *The Non-native Teacher.* London: Macmillan.

Mee-ling, L. (1999). Jet and NET: A comparison of native-speaking English teachers schemes in Japan and Hong Kong. *Language, Culture and Curriculum,* 12, (3), 251-227.

McLaughlin, B. (1987). *Theories of Second-language Learning.* London: Edward Arnold.

Milambiling, J. (2000). Comments on Vivian Cook's "Going beyond the native speaker in language teaching." How nonnative speakers of English fit into the equation. *TESOL Quarterly,* 34, (2) 311-332.

Morgan, G.A., Griego, O.V and Gloeckner, G.W. (2001). *SPSS for Windows: An Introduction to Use and Interpretation in Research.* Mahwah, NJ and London: Lawrence Erlbaum Associates.

Olivares, I. P. (1998). *Oral Proficiency Testing and Language Progress over Time.* Michigan: UMI Dissertation Services.

Oller, J.W. (1976). A program for language testing research. In H.D. Brown (Ed.) *Papers in Second Language Acquisition.* Language Learning, pp. 141-166.

Paikeday, T.M. (1985). *The Native Speaker is Dead.* Toronto and New York: Paikeday Publishing Inc.

Pallant, J. (2001). *SPSS Survival Manual.* Buckingham, UK: Open University Press.

Paterson, L. (1991). An introduction to multi-level modelling. In S. W. Rodenbush and J. D. Willms (Eds.), *Schools Classrooms and Pupils; International Studies of Schooling from a Multilevel Perspective.* San Diego, CA: Academic Press.

Paterson, L. and Goldstein, H. (1991). New statistical methods for analyzing social structures: an introduction to multilevel models. *British Educational Research journal,* Volume 17 (4), pp. 387-395.

Pennycook, A. (1990). Towards a critical applied linguistics for the 1990s. *Issues in Applied Linguistics*, 1, (1) 8-28.

Phillipson, R. (1992a). ELT: The native speaker's burden? *English Language Teaching Journal* 46, (1), 12-18.

Phillipson, R. (1992b). *Linguistic Imperialism.* Oxford: Oxford University Press.

Poon, A.Y.T. (2000). *Medium of Instruction in Hong Kong: Policy and Practice.* Maryland: University Press of America.

Quirk, R. (1985). The English language in a global context. In R. Quirk & H.G. Widdowson (Eds.) *English in the World: Teaching and Learning the Language and Literatures.* Cambridge: Cambridge University Press.

Rampton, M.B.H. (1990). Displacing the 'native speaker': Expertise, affiliation and inheritance. *English Language Teaching Journal* 44, (2), 97-101.

Rasbash, J. et al. (2000). *A Users Guide to MLwiN.* London: Institute of Education.

Richards, J., Platt, J. and Weber, H. (1985). *Longman Dictionary of Applied Linguistics.* London: Longman.

Rice, N. and Leyland, A. (1996). Multilevel models: Applications to health data. *Journal of Health Services Research and Policy.* Volume 1 (3), pp. 154-164.

Samimy, K.K. and Brutt-Griffler, J. (1999). To be a native or non-native speaker: perceptions of 'non-native' students in a graduate TESOL program. In G. Braine Ed.) *Non-native Educators in English Language Teaching.* Mahwah, N.J.: Lawrence Erlbaum Associates, pp. 127-158.

Scully, E. (2001). *Working as a Foreign English Teacher in Rural Japan: JET Instructors in Shimane Prefecture.* ERIC database ED422725.

Shum, H.M. (2001). Perceptions of school culture in Hong Kong: NETs vis-à-vis students. (Un-published M.Ed. dissertation) Hong Kong: Hong Kong University.

Skehan, P. (1998). *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

SPSS Inc. (2001). SPSS for Windows.

Sridhar, K.K. and Sridhar, S.N. (1986). Bridging the paradigm gap: second language acquisition theory and indigenized varieties of English. *World Englishes*, 5, (1), 3-14.

Sridhar, S.N. (1994). A reality check for SLA theories. *TESOL Quarterly,* 28, 800-805.

Stern, H.H. (1983). *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press.

MENETS. (1999). *Monitoring and Evaluation of the Native-speaking English Teacher Scheme (MENETS): First Interim Report.* Hong Kong: Hong Kong Institute of Education.

MENETS. (2001). *Monitoring and Evaluation of the Native-speaking English Teacher Scheme.* Hong Kong: Hong Kong Institute of Education. [unpublished].

Sweeting, A.E. (1990). *Education in Hong Kong Pre-1841 to 1941: Fact and Opinion.* Hong Kong: Hong Kong University Press.

Sweeting, A.E. (1992). Hong Kong education within historical processes. In G.A. Postiglione (Ed.) *Education and Society in Hong Kong: Toward One Country and Two Systems.* Hong Kong: Hong Kong University Press.

Sweeting, A.E. (1998). Education and development in Hong Kong. In P. Stimpson and P. Morris (Eds.) *Curriculum and Assessment for Hong Kong.* Hong Kong: Open University of Hong Kong Press.

Tabachnick, B. G. and Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd edition). New York: Harper Collins.

Tajino, A. and Tajino, Y. (2000). Native and non-native: What can they offer? Lessons from team teaching in Japan. *ELT Journal,* 54, (1) 3-11.

Tang, C. (1997). The identity of the nonnative ESL teacher. On the power and status of nonnative ESL speakers. *TESOL Quarterly.* 31, (3) 577-579.

Tay, M. (1992). The uses, users and features of English in Singapore. In J. Pride (ed.) *New Englishes.* Rowley, MA: Newbury House.

Thomas, J. (1999). Voices from the periphery: non-native teachers and issues of credibility. In G. Braine Ed.) *Non-native Educators in English Language Teaching.* Mahwah, N.J.: Lawrence Erlbaum Associates, pp. 127-158.

Tung Chee Hua. (1997). Policy address speech 1997. (Full text available from the official government Web Site: 'Hong Kong Special Administrative Region of the People's Republic of China.' http://www.policyaddress.gov.hk/pa01/eindex.html

Tung Chee Hua. (2001). Policy address speech 2001. (Full text available from the official government Web Site: 'Hong Kong Special Administrative Region of the People's Republic of China.') http://www.policyaddress.gov.hk/pa01/eindex.html

Tymms, P. (1997). *The Value Added National Project. Technical Report: Primary 4.* Durham: Curriculum, Evaluation and Management Centre, University of Durham.

Ur, P. (1981). *Discussions that Work.* Cambridge: Cambridge University Press.

Van Lier, V. (1989). Reeling, writhing, drawling, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly,* 23 (3), 489-508.

Walker, E. (2001). Roles of native-speaker English teachers (NETs) in Hong Kong secondary schools. *Asia pacific Journal of Language in Education. 4* (2).

Watcyn-Jones, P. (1981). *Pair Work: Activities for Effective Communication.* Harmondsworth, UK: Penguin Books.

Weir, C.J. (1990). *Communicative Language Testing.* New York and London: Prentice Hall.

Wolfe, E.W., and Chiu, C.W.T. (1999a). Measuring Pretest-postest change with a Rasch rating scale model. *Journal of Outcome Measurement, 3,* 134-161.

Wolfe, E.W., and Chiu, C.W.T. (1999b). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement, 3(4),* 360-381.

Wright, B.D., and Masters, G.N. (1982). *Rating Scale Analysis.* Chicago, IL: MESA Press.

Wright, B.D., and Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1*(1), 83-106.

APPENDIX I

CONFIDENTIAL

Marking Criteria for Secondary Oral Assessment

Level	Comprehension/ Communication	Fluency/ Productivity	Vocabulary	Grammar	Pronun-ciation
6	Pupil communicates with relative ease on a range of topics and understands longer stretches of connected discourse at normal rate.	Pupils converses with increasing fluency, spontaneity and creativity; initiates, elaborates and sustains connected talk.	Pupil has a broad enough vocabulary for discussing simple social and general interest topics.	Uses simple tenses but lacks full control of more complex forms. Occasional grammatical inaccuracies persist.	Little L1 interference in sounds, stress and intonation. Communication not impaired.
5	Pupil can communicate and understand new sentence-level questions and statements on a good range of topics at normal rate of speech. Communication breakdown is rare.	Speech, while hesitant, is gaining in coherence, speed and length but pauses still noticeable. Attempts to use longer, more complex sentences.	Pupil has adequate vocabulary for sentence-level conversations on wider personal topics with gaps when speaking on general interest topics.	Maintains simple conversations, mostly in present tense with some errors; speech has some grammatical inaccuracies.	Some L1 interference in sounds, stress and intonation evident but unobtrusive. Communication seldom impaired.
4	Pupil is able to communicate and understands new sentence-level questions and statements on a limited range of topics at normal rate of speech. Occasional breakdown of communication may occur.	Speech less hesitant. pupil produces simple sentence level conversation with success, and some ease in everyday interactions.	Vocabulary adequate for simple questions and statements to satisfy basic needs but not sufficient for explanation or elaboration.	Creative language use at sentence-level; verbs conjugated but mostly inaccurate; speech has many grammatical mistakes.	L1 interference in sounds, stress and intonation noticeable and may be obtrusive but generally comprehensible to listener with occasional strain.
3	With some repetition and rephrasing, pupil is able to communicate and understand simple questions and statements on familiar topics at slower than normal rate of speech.	Speech still hesitant but produces memorised expressions with ease. Pupil makes limited attempt at creative sentence formation.	Vocabulary for basic objects, places and kinship terms and utterances on predictable topics.	Memorised expressions used accurately and with ease. Some creative sentence-level speech with verbs generally lacking or unconjugated.	Fairly strong L1 interference in sounds, stress and intonation causes some misunderstandings requiring repetition and placing some strain on the listener.
2	Pupil has very limited functional communicative ability but understands some highly predictable utterances in familiar topics at slower than normal rate of speech.	Utterances very hesitant and often incomplete except in a few stock phrases. Sentences, when attempted, are disjointed and restricted in length.	Some specific words and memorised expressions on limited topics. Frequently searches for words.	Memorised expressions used accurately with two/ three-word phrases but with only a few isolated verbs used, mostly not present in created sentences	Strong L1 interference in sounds, stress and intonation causes frequent misunderstandings. Repetition often needed and considerable strain placed on listener.
1	Pupil has essentially no functional communicative ability but recognises isolated words and high frequency expressions.	Speech halting and fragmentary. Produces only isolated words and a few high-frequency expressions.	Restricted to a few isolated words and memorised expressions on a few familiar topic areas.	Memorised, high-frequent phrases used with words juxtaposed at random without consistent knowledge of basic s-v-o sentence structure.	Major problem in sounds, stress and intonation cause serious comprehension problems for listener.

Stage 5 RSM Control File

```
TITLE = "1st Cohort 2nd Admin STEP 5 - Oral Assessment"

;Input Data Format
 NAME1 = 1      ; column of start of person information
NAMLEN = 5      ; maximum length of person information
 ITEM1 = 7      ; column of first item-level response
   NI = 5       ; number of items = test length
 XWIDE = 1      ; number of columns per response
PERSON = student  ; Persons are called ...
 ITEM = Item    ; Items are called ...

DATA = 12secdf.prn

pafile=12secSTEP4opf.txt
safile=1secSTEP2osf.txt


ifile=12secSTEP5oif.txt
pfile=12secSTEP5opf.txt
sfile=12secSTEP5osf.txt

; For rescoring
;     0      1      2      3      4      5      6      7
;     12345678901234567890123456789012345678901234567890123456789012345678901234567890
;GROUPS=0       ; specify that each item has its own rating scale (partial credit)
;IREFER=AAAAAAAAAAAAABBBBBBBBBBBBBBBBBBBBBBBBBCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCC

; Data Scoring
  CODES = "123456"    ; valid response codes
;IVALUEA= "01"   ; for rescoring for item type A
;IVALUEB= "10"   ; for rescoring for item type B
;IVALUEC= " "    ; for rescoring for item type C
;             ; Codes in IREFER with no IVALUE are not changed

CFILE = *           ; label the categories in Table 3
3+0 Strongly Disagree ; 3+0 means item 3 (an example of its group),
3+1 Strongly Agree    ; 0 is the value after any rescoring, keying, etc.
*

;NEWSCORE = "10"    ; use to rescore all items
;   KEY1 =          ; key for MCQ items

; XWIDE = 2   ; for all codes 00 to 99
; CODES = "000102030405060708091011121314151617181920212223242526272829+
;       +303132333435363738394041424344454647484950515253545556575859+
;       +606162636465666768697071727374757677787980818283848586878889+
;       +909192939495969798 99"

; codes reversed, in case needed
; NEWSCR= "999897969594939291908988878685848382818079787776757473727170+
;       +696867666564636261605958575655545352515049484746454443424140+
;       +393837363534333231302928272625242322212019181716151413121110+
;       +09080706050403020100"

MISSING = 1

;User Scaling
```

```
UMEAN  = 0          ; item mean - default is 0.00
USCALE = 1          ; measure units - default is 1.00
UDECIM = 2          ; reported decimal places - default is 2
MRANGE = 0          ; half-range on maps - default is 0 (auto-scaled)

&END

;Put item labels here for NI= lines
comprehension
fluency
vocabulary
grammar
pronunciation
END LABELS
;Put data here
```