EVALUATION REPORT

EV485

THE ROLE AND DESIGN OF BASELINE STUDIES IN THE EVALUATION OF ENGLISH LANGUAGE TRAINING IN THE CASE OF NEPAL

BY

JOHN BORTON (ODA)
DR CYRIL WEIR
JOHN ROBERTS

CONTENTS

ANNEXES

PREFACE

Each year the Overseas Development Administration (ODA) commissions a number of ex post evaluation studies. The purpose of the ODA's evaluation programme is to examine rigorously the implementation and impact of selected past projects and to generate the lessons learned from them so that these can be applied to current and future projects.

The ODA's Evaluation Department (EVD) is independent of ODA's spending divisions and reports direct to the ODA's Principal Finance Officer.

Evaluation teams consist of an appropriate blend of specialist skills and are normally made up of a mixture of in-house staff, who are fully conversant with ODA's procedures, and independent external consultants, who bring a fresh perspective to the subject matter. This particular evaluation report was completed by an in-house member of staff, and concerns a baseline study completed by outside consultants.

The baseline study involved the following stages -

- appointment by EVD of an independent consultant to make an initial desk study of all relevant papers and a preparatory visit to Nepal;

- visits by the UK freelance consultants from the University of Reading to establish and monitor the procedures by which local consultants would gather data;

- data analysis in the UK and production of reports by the consultants;

The evaluation study involved -

- production of a first draft report which was circulated for detailed comment to the individuals and organisations most closely concerned;

- submission of the revised draft report to the ODA Principal Finance Officer, to agree

the main conclusions and lessons to be learned from the study, on the basis of the draft report;

- completion of this final report which is published along with an Evaluation Summary (EVSUM).

This process is designed to ensure the production of a high quality report which draws out all the lessons.

J C H Morris,

Head, Evaluation Department.

ABBREVIATIONS

ADB	Asian Development Bank
BC	British Council
CSTDC	Curriculum Supervision Textbook Department Centre (Nepal)
Control Group	The group of students of teachers who were not trained under the project.
ELT	English Language Teaching
Experimental Groupwere	The sample group of students of teachers who trained under the project.
EVD	Evaluation Department ODA
HMGN	His Majesty's Government of Nepal
INSET	In-service Education of Teachers
KELT	Key English Language Teaching
ODA	Overseas Development Administration
SEADD	South East Asia Development Division of ODA
SEP	Science Education Project
SEPELT	Science Education Project English Language Teaching
SLC	School Leaving Certificate
TC	Technical Cooperation
UNDP	United Nations Development Programme

OVERVIEW

The Study

1. Evaluation Department has undertaken two studies on the effectiveness of English Language Teacher Training, the first in Nepal and the second in Guinea. As initially conceived, the Nepal study set out to generate a method for the interim and ex post evaluation of ELT projects by establishing the nature of the baseline data required for such evaluation; to develop methodologies for the collection of such data; and to collect these data. The study was thus a vehicle for the development of methods and methodologies, with the subsequent Guinea study serving to test and refine these outputs of the Nepal study. Each study, therefore, generated lessons both for the design and implementation of baseline studies in future, and for the design and implementation of ELTT projects.

2. This report concerns what was originally to be a baseline study of a secondary teacher training project in English Language Teaching in Nepal. The study was undertaken by UK freelance consultants from the University of Reading, with assistance from local consultants, New Era, in Nepal. It was funded by the Evaluation Department (EVD) of ODA. The study took place during 1989 and 1990, with a final report being produced in 1991.

3. The methodology of the study was to identify and collect baseline data that would illumine the impact of the teacher training project on the performance of the students. This was done through a series of language achievement tests conducted on three occasions (March and November 1989, and November 1990). The tests were administered both to an "experimental" group of students of teachers who had been trained under the project and, for comparative purposes, to a "control" group of students whose teachers had not been trained. Further data were gathered on the teachers and the schools in order to isolate the impact of the teacher training, using computerized statistical analysis.

4. It was expected that the baseline study would be an input into an ex post evaluation of the project. This was not possible, however, due to the change both in the project and in the control group of the baseline study; what evolved, therefore, was a study which also sought to shed some light on the impact of the project. This evaluation report summarises the findings of the baseline study which were presented in various reports by the consultants, and adds an analysis of the use of baseline studies for future projects of this type.

5. The study has shown that it is feasible to mount a baseline study, even in a country with very limited communications networks such as Nepal. From the results obtained, the procedures developed in the study would appear for the most part to be effective, and should be transferable to other similar projects. They would be cheaper if conducted by project staff, possibly with limited external supervision. The development of language tests to measure the impact of teacher training programmes on student language performance may mark the first real step towards the design of reliable and valid test instruments in the country concerned.

The Secondary Teacher Training Project

6. The principal objective of the project that the study examined was the provision of an intensive one month teacher training course in English Language Teaching to 900 secondary school teachers. This was achieved by the provision of one Key English Language Teaching (KELT) officer for a period of two years, who provided training to local teacher trainers and prepared a training manual. The teacher training courses were then implemented by the local trainers with support from the KELT officer as necessary.

7. The overall cost of the project was ,313,000, with a unit cost per teacher trained of ,346. This is cost-effective, but the project had not reached sustainability at its conclusion.

Overall Conclusion of the Study

8. The overall conclusion of the study was that the project had a small but statistically significant impact on the performance of the students of the trained teachers compared with that of the students of untrained teachers. The improvement in the performance was more marked in the first year of the study but performance continued to improve, at a slower rate, into the second year.

9. When considering subgroups of "high", "medium" and "low" ability students, the only significant gains were achieved by students who were of a high ability at the start of the study. This suggested that students who had almost no ability in English at the start of the study were not able to benefit from the improved teaching practices that the training encouraged.

10. Observations of teachers showed very significant differences between the control and experimental group in terms of teaching practice adopted. Testing of teachers revealed some small but significant differences in oral and grammatical ability, in favour of the trained teachers.

Main Findings of the Study

11. A full summary of findings is provided in Chapter 3.5. and provided in more detail in the Annex. The statistically significant findings are summarized here.

a. Over the whole period of the study, the experimental group's average score improved by 17 percentage points in a dictation test, compared with an increase of 7 points in the control group.

b. During the second year of the study alone, the average dictation score of the "high" ability students in the experimental group improved by 17 percentage points, whereas the performance of the "high" ability students in the control group declined.

c. Over the whole period of the study, the "high" ability experimental sub group's average score in the dictation test improved by 40 percentage points, compared with only a one percentage point improvement for the "high" ability control sub group.

d. Over the whole period of the study, the "high" ability experimental subgroup's average score in a comprehension ("gap filling") test improved by 15 percentage points, whilst that of the "high" ability control subgroup declined slightly.

e. Observational data revealed that trained teachers scored an average of 67% in

terms of using teaching practices defined as characteristic of a trained teacher, compared with an average score of 24% for untrained teachers.

Lessons Learned

Appraisal

Lesson 1. Baseline studies should ideally be appraised at the same time as the project they are assessing. A baseline design should be incorporated in the overall project design from the start.

Lesson 2. As well as helping to clarify the objectives for those concerned, the design of the study instruments could facilitate the monitoring of national standards in language performance, particularly where the national public examinations are inappropriate.

Design

Lesson 3. Education baseline studies require particularly large amounts of staff time in order to collect and process the data.

Lesson 4. Care must be taken in choosing appropriate test instruments for educational baseline studies, and the instruments must relate to the objectives of the teaching if they are to demonstrate the impact of the project. In general it is advisable to have a range of tests, as some tests may not provide useful results.

Lesson 5. Teacher observations can help project staff to realise training objectives in precise terms, can help to explain the student test results, provide more confidence in the conclusions of an educational baseline study, and provide valuable information on the implementation of training in the classroom.

Lesson 6. Adjusting for other variables which might affect the student test results such as size of class and number of hours taught also provides more confidence in the conclusions of the data analysis.

Lesson 7. The size of the sample of teachers to be included in an educational baseline study must take account of likely attrition during the study, and the longer the period of study, the larger the initial group must be.

Lesson 8. Evaluations based on educational baseline studies will be most meaningful if they are conducted over an extended period of time, and include post-education monitoring. Only projects planned to last long enough to identify project outcomes fairly should incorporate baseline studies.

Lesson 9. The methodology of data analysis and format for reporting should be clearly established before the start of an educational baseline study, and interim reporting kept to the minimum level necessary for monitoring purposes.

Staffing

Lesson 10. The baseline study design may take a number of forms according to host country conditions. The main staffing options are local staff, local consultants, expatriate project staff, and expatriate consultants.

Lesson 11. External consultants are justified for those education projects where commensurate benefits to the education system or project planning are expected to accrue. This means that only longer and bigger projects or pilot projects which, by definition, are likely, if successful, to be expanded in the future, would usually justify an externally validated baseline study.

Lesson 12. Even when not undertaking the primary data collection and analysis, an external consultant can be useful for validation and interpretation at agreed 'milestones' in the project.

Lesson 13. Where project staff are used for a baseline study, their job description(s) should include the systematic collection of implementation data according to a specified framework of content and method appropriate to monitoring and evaluation.

Lesson 14. Education baseline studies may be more cost-effective if they use local consultants (or counterpart local staff), where it is possible to find suitable individuals, and this may be the only justifiable way to implement a baseline study of a small project. The use of local staff would have training benefits which may contribute to sustainability. Training and supervision of such staff may be necessary in order to ensure that accurate data are collected. With this solution, monitoring requirements should be written into the project framework and job descriptions of teacher trainers.

Lesson 15. Close cooperation is needed between those administering the baseline study, and those administering the project. Liaison and agreement between ODA's local representative and the government is necessary at the outset, with a clear statement of the respective responsibilities of all parties.

Usefulness of Results

Lesson 16. Unless an education baseline study attempts a "needs analysis" to define a standard of education required for the labour market and education system in a particular country, and monitors performance of students against that standard, there is no bench mark against which to assess the results of the baseline study.

Lesson 17. Without such an analysis, an education baseline study might be useful for comparing the relative efficacy or educational impact of different types of intervention, but is not particularly useful in assessing the impact of a single intervention.

Lesson 18. The value of educational baseline studies will be greater than was the case here, if they are conducted for continuing projects, where the outcome of the baseline study is likely to be of use in future project design.

Design of ELT Projects

Lesson 19. The immediate objectives of ELT projects should be specified in terms of the changes anticipated in teaching practice and student performance; and, where a baseline study is conducted, it should gather data relevant to these objectives.

Lesson 20. ELT project appraisals should examine the suitability of the formal examination system, as training geared at skills not included in examinations is unlikely to be transferred to the classroom.

Lesson 21. Detailed training design should be appraised taking account of both the unit costs

and the likely impact on quality of training.

Lesson 22. Sustainable teacher training projects will require the teacher training team to be formally appointed as trainers under an appropriate institution. Ad hoc trainers are unlikely to command the resources necessary to sustain the training programme.

Lesson 23. ELT projects should be based on a detailed manpower needs analysis, in order to develop meaningful wider objectives.

MAIN REPORT

1. BACKGROUND

1.1 In 1988 a baseline study was commissioned by ODA's Evaluation Department (EVD), of an ODA-funded English Language Teaching (ELT) project in Nepal.

1.2 The project that the baseline study examined was the ELT teacher training component of a wider in-service teacher training Science Education Project (SEP) which was jointly funded with the Asian Development Bank (ADB), United Nations Development Programme (UNDP), and His Majesty's Government of Nepal (HMGN). The ODA-funded component lasted for two years, 1987 - 1989, with some follow-up assistance which ended in 1991.

1.3 The SEPELT (Science Education Project English Language Teaching) INSET (In-service Education of Teachers) scheme in Nepal was set up to provide 1080 grade 8 - 10 (upper secondary level) English Teachers with one month's in-service training delivered by locally-trained Nepali staff, working from a standard course manual and supported by an expatriate training officer, funded by ODA under the Key English Language Teacher (KELT) scheme.

1.4 It had been envisaged that the baseline study would be used as an input into a full evaluation of the project. In 1992 it was decided not to pursue an evaluation, because the validity of the baseline study's "control group" of teachers not benefiting from the INSET training had been largely destroyed by the project's decision to offer training to most of the group's members.

1.5 It was, however, decided to make an evaluation of the baseline study, for the following reasons:-

a. The ODA had not previously made use of a baseline study in its education sector evaluations, and it was considered important to disseminate the results for use by those considering the adoption of similar techniques for monitoring and evaluation of projects of this type;

b. Although the various consultants' reports contain most of the information presented here, it was considered useful to present the results in a single report, and to provide an assessment of the study from the perspective of the commissioning organization;

c. The baseline study represents a considerable investment for EVD, and it is appropriate that commensurate efforts are devoted to the feedback of this information in the form of a report that can be distributed to interested parties.

1.6 Preparatory work to establish the baseline study was undertaken for EVD by an independent consultant, Mr Brook. The study itself was made by consultants from the University of Reading. The team was led by Dr C. Weir, with support from Mr J Roberts, particularly in the observational data collection and analysis, and to a minor extent, by Mr Hughes. Data collection in Nepal was undertaken by a local consultancy firm, New Era. ODA management of the baseline study was provided by EvD economic advisers.

1.7 This evaluation was made by Mr J Burton, Economic Adviser, EVD, and the University of Reading Consultants, Dr C Weir and Mr J Roberts, on the basis of the reports (listed at Appendix 1) of those who undertook the study and other material on file. In order to avoid extensive use of quotation marks in this report, some text from the various baseline study documents has been used directly in this report without detailed attribution.

1.8 This report first contains in Chapter 2 a summary of the project that the baseline study was examining. This is essential in order to understand the significance of the baseline study. Chapter 3 is concerned with the baseline study itself, and presents a summary of the results of the study, (which are provided in more detail in the Annex), before drawing conclusions about the usefulness of such studies.

2. THE SECONDARY EDUCATION PROJECT ENGLISH LANGUAGE TEACHING

2.1 English in the Secondary Education System in Nepal

2.1.1 As the focus of this report is upon the results of the baseline study, it is not necessary to provide an extensive analysis of the education system in Nepal. A brief summary, however, of the scale and structure of the secondary education sector is necessary as background material.

2.1.2 In 1988 there were 1400 Upper Secondary Schools in Nepal, with around 8,000 teachers, and 242,000 pupils. Enrolment at that level was 21% of the total age group, compared with an enrolment rate of over 80% at primary level.

2.1.3 English teaching commences in grade 4. According to a survey in 1983 carried out by ODA (the Davis Report), pupils appeared to learn almost no English in the first four years (a finding confirmed by the baseline study), and it was recommended that English be deferred until grade 8 given the lack of resources. This recommendation was not implemented.

2.1.4 The School Leaving Certificate (SLC) examination is completed in grade 10. There was a screening test for SLC, administered by the District Education Office. Around 25% of those entering the screening test were not allowed to enter the SLC and, of those taking the SLC, typically around 65% failed. English was one of the compulsory subjects in the SLC examination, and was also one of the two subjects (the other being mathematics) which caused most failures of the SLC.

2.2 Identification, Design and Appraisal

2.2.1 The origin of the SEPELT project is somewhat obscure. The 1983 Davis report

recommended that in-service education should be provided for all teachers of English, (amongst others). The decision to focus on in-service education of teachers, however, (which was only one of the Davis report recommendations) is not documented in ODA.

2.2.2 A follow-up visit in March 1985 produced a report entitled "A Project Design for the In-Service Education of Upper Secondary School Teachers of English in Nepal". This recommended the appointment of a KELT officer to the Curriculum Supervision Textbook Department Centre (CSTDC). During 1986 a series of four two week ELT seminars were arranged.

2.2.3 In the light of these seminars the job description for the KELT officer was revised. The officer provided teacher training to local trainers, who in turn delivered one month teacher training courses. The teacher trainers were not specialist trainers, but were drawn from the teacher and headmaster population.

2.2.4 The link with the ADB / UNDP / HMGN Science Education Project appeared to arise informally out of discussions between the British Council (BC) Representative and the Ministry of Education about the workplace of the proposed KELT officer. The Ministry suggested the SEP building, and this was accepted by ODA. The location of the KELT officer with the CSTDC was therefore dropped. The UNDP was initially reluctant to accept the ELT component in its project.

2.2.5 Mr Brook (March 1988) stated that there had apparently been no assessment of these institutional arrangements in the light of alternatives and there was a lack of an ELT needs analysis prior to the appointment of the KELT officer in 1987. It was not clear who actually used English, and how the supply of English skills was deficient in meeting the demands of the school and higher education system and the labour market.

2.2.6 A further problem highlighted (May 1988) was that the examination system required reform, as it tended to encourage learning by rote rather than English as a medium of communication. In the absence of such a reform, Mr Brook questioned the sustainability of a project which encouraged innovative language teaching methods.

2.2.7 More fundamentally, given that the objective of the project was to improve performance in the SLC and given that the University (which accepted all students with suitable grades) was already over-stretched, it was not clear that the project would have useful results. The short term outcome would be to exacerbate the University's resource problem and the long term result might be an increase in the incidence of educated unemployment, which was already a major problem.

2.2.8 Given that over 90% of the population was involved in agriculture and living in highly remote areas without need for English, Mr Brook considered that the attempt to improve skills universally was highly questionable. Whilst English was undoubtedly required at University Level, a more focused approach should be adopted to address this need.

2.3 Inputs and Costs

2.3.1 Details of the ODA project costs are provided at Annex Table 1. Excluding UK training awards, the total project cost was nearly ,203,000. The main input financed from this amount was the provision of the KELT officer. The KELT scheme provides fully funded technical co-operation staff which are administered by the BC. The officer was appointed in 1987 for a

period of two years.

2.3.2 The project also provided support costs such as a vehicle, and funding of a project assistant. A short ELT follow-up consultancy was provided in 1990. TC training awards were provided in support of the project (details of which are provided at Annex Table 2). 18 awards were provided over 1988/89 - 1992/92 at a cost of ,111,000.

2.3.3 Including UK training, the total cost of the project was slightly in excess of ,313,000. The scale of the ODA contribution was a relatively small proportion of the total SEP, with a budget of $13 million.

2.4 Objectives

2.4.1 The long-term goal of the training was the improvement of student performance in the SLC English Examination, through assisting pupils in grades 8, 9 and 10 of secondary schools in all regions of Nepal. The specific objectives included:

a. identifying the English language needs of Upper Secondary English pupils and teachers;

b. developing, running and supervising in-service English language teaching courses which in form and content optimally meet the identified needs;

c. identifying and orientating personnel for English language in-service courses in all regions of Nepal.

2.4.2 The courses provided training to teachers in basic ELT procedures, with a view to enhancing the teaching of the National English Curriculum. Teachers' language improvement was a supplementary objective of the courses.

2.4.3 The initial request from HMGN suggested that the project would involve 450 teachers in a 2 month training period. By the time the training commenced this had been adjusted by a decision of the Ministry of Education, to 1020 teachers to be provided with one month of training.

2.4.4 The KELT reported in May 1989 that the target had been reduced to 900 teachers, and it also had been decided that follow-up training of a further two weeks should be provided for all 900 of the teachers trained. In July 1990 the BC reported that the target for the follow-up training had been reduced to 300 teachers, as it was not possible to train 900.

2.5 Other Inputs Considered

2.5.1 Other inputs considered but not implemented were the appointment of two additional KELTS, the renewal of the existing KELT post (with a new officer, as the incumbent was not available for a second contract), and a second short consultancy visit.

2.5.2 In November 1987 the ODA Education Adviser recommended an additional two KELT posts, one in curriculum development and one in teacher training. These posts were agreed by the ODA in April 1988, and a draft project memorandum and project framework for all three posts were drawn up.

2.5.3 In August 1988 the Ministry of Education requested that these two posts be withdrawn due

to uncertainties that the SEP would continue, although it requested that a second contract be offered for the existing KELT post. The ODA Education Adviser, in January 1989, recommended the ending of the ODA project until it was clear that a second phase of the SEP would proceed.

2.5.4 Although in May 1990 the BC reported that the ADB funding of the SEP was to be extended to June 1991, the KELT officer was not replaced following his departure in June 1989. Instead the BC proposed that support be provided through short consultancy visits. Two such visits were approved by ODA in September 1990, but only the first visit took place, in December 1990. The second did not take place because approval of the visit was conditional on HMGN meeting all local costs, which in the event was not agreed.

2.6 Outputs

2.6.1 The project was successful in meeting the revised target of training 900 teachers for a period of one month. In fact 903 were trained by June 1990. This output was achieved through the training of 26 teacher trainers under the project.

2.6.2 The project was less successful in providing the two weeks follow-up training, even for the revised target of 300 of the teachers who had attended the one month course. The December 1990 consultancy visit report stated that only 70 teachers had been provided with follow-up training.

2.7 Conclusions

2.7.1 Although this report does not attempt to provide an evaluation of the SEPELT project, it is useful to at least draw conclusions on the available information.

2.7.2 The training of 903 teachers at a cost of ,313,000 for a period of 4 weeks in-service teacher training for all teachers, and 6 weeks for 70 teachers, suggests a unit cost of around ,15 per training day. This appears to be a cost-effective delivery of training, and it is arguably probable that the benefits will have exceeded the project costs, given the changes in teaching practices revealed in Chapter 3, and the Annex.

2.7.3 The lessons that might be drawn from the project are largely already learned by ODA. In July 1988 a Policy Guidance Note was issued on English Language Teaching (i.e. after the design of this particular project). Amongst the guidance are the following lessons:-

a. "the particular focus of an ELT project should take account of economic and manpower needs;"

b. "it should always be formulated in a project format, including any combination of manpower, training, equipment, books, capital expenditure and institutional links;"

c. "in the design of an ELT project careful attention must be given to the localisation phase and to the project's sustainability by the host country after ODA support has ceased."

2.7.4 It is clear from the comments of Mr Brook that the SEPELT was not based on a detailed manpower needs analysis, and therefore may have had little economic benefit, because, for example, the majority of pupils do not go on to higher education and have no economic need for English language.

2.7.5 Although the project did include manpower training, some of it overseas, relatively little was done to supply necessary equipment and books to create a sustainable training capacity. ODA and HMGN also appear to have failed to agree a clear set of detailed objectives and the associated inputs. Section 2.5 is evidence of a lack of certainty over inputs.

2.7.6 The switch from the planned two month courses to one month courses (para 2.4.3) is also relevant. Detailed training design should be appraised taking account of both the unit costs and the likely impact on quality of training.

2.7.7 The sustainability of the project appears to have been affected by the lack of effective institutional arrangements for localisation. The teacher training team was formed from teaching staff who were not formally appointed as trainers, and who returned to their original posts after the project.

2.7.8 Although the project made some attempt to assess the training through questionnaires to trainees, there was no formal follow up to assess the impact of the training.

2.7.9 The comments at para 2.2.6 suggest that the impact of training would be limited by the inappropriate nature of the examination system. This suggests that for the project to be fully effective it would have been necessary to reform the examinations.

3 THE STUDY

3.1 Background

3.1.1 The decision to undertake a baseline study as a contribution to an evaluation of the impact of ELT Projects was initiated by the Education Division of ODA, which requested EVD to assist in developing baseline studies of a number of ELT projects. In 1988 a consultant, Mr Brook, was commissioned by EVD to examine the suitability of various ELT projects.

3.1.2 His March 1988 report established the methodology for this baseline study. The immediate objective was to measure the impact of the SEPELT teacher training on the English language performance of students. Details of the methodology are provided in section 3.3.

3.2 Inputs and Costs

3.2.1 The initial costs incurred in setting up the study included a preliminary visit to Nepal in May 1988. The total cost of this element was nearly ,4,150.

3.2.2 At appraisal (i.e. the initial visit report), there was no detailed estimate of costs, although it is clear that they were envisaged as being much less than the actual cost. It was expected that the project manager would be a KELT officer. The costs of the project manager were estimated at ,4,000 in the first year for two-man months, with only one man-month in the two subsequent years, plus sundry costs of ,2,000. It was anticipated that local consultants would be engaged for data collection but the cost of this was not estimated.

3.2.3 The method of implementation had to be changed, however, because the Ministry of Education decided not to request the additional KELT appointments that had been envisaged (see paragraph 0). It was therefore necessary to use a mixture of external and local consultants. A draft project framework was prepared in June 1988 which estimated the cost (without a

detailed implementation plan) at ,20,000. EVD contracted consultants to assess feasibility of the study in June 1988.

3.2.4 The principal input was the UK consultancy services. This comprised a total of 165.5 man days. Dr Weir's input amounted to 106 days, Mr Roberts spent 53.5 days on the study, and Mr Hughes 6 days. They made a total of three visits to Nepal:-

a. In November 1988 Dr Weir and Mr Roberts visited to agree details of the study with KELT and negotiate a contract with New Era;

b. In January 1989 Dr Weir visited to train New Era Staff in observation techniques and to administer tests; and

c. In November 1989 Dr Weir and Mr Roberts visited to monitor the second administration of tests and observations.

3.2.5 The project also allocated funds for a Nepalese consulting firm "New Era" to administer language tests to students and observe teaching methodology. The total input from New Era was 330 days, at a cost of just under ,4,000.

3.2.6 The total cost was in excess of ,42,000. This amount was spread over four financial years, and covered the UK preparatory consultancy, the main UK consultancy, and the local consultancy by New Era, as detailed in Annex Table 3.

3.3 Methodology of the Baseline Study

Implementation Arrangements

3.3.1 At appraisal a number of options for implementing the baseline study were identified. These included:

a. A KELT Project Manager. It was recommended that one of the expected two additional KELT officers should take responsibility for the baseline study.

b. Use of Local Governmental Institutions. This was rejected because of a lack of institutional capacity.

c. Local Consultants. The use of New Era consultants was identified as a possibility, but not recommended because of their lack of experience with language programmes.

d. Outside Consultants. It was recommended that outside consultants be used for designing the test instruments.

3.3.2 For reasons outlined in 3.2 above, it was decided that outside consultants from the University of Reading would be engaged. After their initial consultancy visit it was decided that the use of local consultants should be pursued. The principal reasons for this, as reported in the November 1988 consultants report were as follows:-

a. the experience of the firm in research, including a previous ODA contract;

b. the independent status of the firm which made it a more suitable choice than using

project staff;

c. the likelihood of higher motivation of a consultant compared with a member of the recipient government department, who would have no specific financial incentive.

An additional attraction of New Era was the very cost- effective service provided. New Era also had systems for efficient monitoring of individual staff performance.

3.3.3 New Era staff were trained by the consultants during the second visit in January 1989, and a training manual was prepared. Joint observation sessions by the UK and local consultants showed that they had been trained effectively.

3.3.4 Although the use of TC staff for a baseline study would have been desirable, and would have allowed a reduction in the amount of visiting required by the consultants, the existing project KELT did not have the time to do this given the demands of the project, and was due to leave Nepal before the end of the baseline study. The additional KELT staff were not expected to be in post before the end of 1989, and in fact the request for them did not materialise (see paragraph 0). It was therefore decided to proceed with consultancy supervision.

Approach

3.3.5 The methodology of the Baseline study was largely established following the KELT's initial visit in May 1988. The approach proposed was to measure the success of the ELT project through testing of the teachers being trained, prior to and after training, and the testing of their students immediately after training and at a later date. The details were refined by the consultants during their subsequent visits. The evolution of the methodology of the study is summarized below.

3.3.6 Despite the expression of some reservations about the feasibility of testing teacher performance quantitatively, teacher tests were included in the baseline study (see Annex para 0). The purpose of these tests, however, was not to assess the impact of the training on the performance of teachers, but rather to ensure the equivalence of the control and experimental groups. They were not conducted prior to the training. Pre-training tests would have been useful in assessing the impact of training on teacher performance.

Target Group

3.3.7 The focus of the study on grade 8 students was determined at the preparatory stage of the study. The aim was to study the performance of upper secondary school students whose teachers had been trained under the project, up to their SLC examinations. It was obviously necessary to start monitoring student performance over a period before the SLC if the project impact was to be evident.

3.3.8 It was proposed that tests should establish the comparability between the control and experimental groups, establish the gain in standards over a three year period, and so determine the relative performance of the two groups.

3.3.9 The consultants' first visit report recommended that the selection of control and experimental groups be deliberately structured. For the experimental group the criterion set was that teachers chosen were thought likely to implement their training.

3.3.10 For control and experimental group schools it was recommended that the schools be as close as possible in terms of standards prior to training. This criterion was elaborated in the March 1990 report to comprise:

a. pupil equivalence - established through an equivalence of SLC results; and

b. teacher equivalence - established by asking teachers to complete part I of the students' test in order to assess teachers' language ability.

3.3.11 For both groups, the following additional features were identified:

a. that there should be no special features of the schools that would bias the results;

b. access to the schools by technical staff should be both possible and welcome;

c. the teachers should remain with their grade 8 classes and be likely to continue with the same pupils through grade 9;

d. there should be equivalent stability/ rates of attrition in the control and experimental groups;

e. all schools should be well enough run to ensure efficient data collection;

f. there should be adequate facilities for testing to minimize pupil copying;

g. control teachers should not receive informal secondary training during the study, for example by contact with trained teachers;

h. control teachers should not attend training before late 1990;

i. schools should be easily accessible.

3.3.12 Although it was intended that all these criteria should be taken into account in the selection of teachers for inclusion in the study, the consultants admitted in their March 1990 report that there were information gaps, and the selection was made on the basis of available knowledge.

Sample Sizes

3.3.13 Mr Brook noted (in his visit report) the trade-off between the number of students to be included in the study, and the problem of wastage if a low sample size was selected. The option of sampling within the classes - in other words testing say ten pupils from a given class rather than the whole class - was explicitly mentioned, but this was rejected because of high student wastage. There is no definitive recommendation of the numbers of teachers to be included in each group, but samples of both 30 and 50 (in each group) are mentioned in the report.

3.3.14 Dr Weir recommended in July 1988 that the number of teachers should be reduced to 20 in total because of cost and logistical constraints. Following the consultants' initial visit in November 1988 arrangements were made for New Era to visit 16 experimental and 16 control schools in January 1989 with a view to identifying those suitable.

3.3.15 The consultants stated that if statistical random sampling were to be used, it would be necessary to select three hundred teachers out of the total population, which would be

prohibitively expensive. They recognised that a sample of 10/12 teachers in each group would not constitute a "true experimental design", but believed that the results of a small sample would still be meaningful.

3.3.16 During the initial visit in January 1989, the sample group was scaled down from the 32 teachers initially selected to 24 - 12 in each group. This was done on the basis of the tests of teachers' language ability (see paragraph 0). This was reduced through attrition of two teachers between February and November 1989 to 11 teachers in each group, by the time of the second test battery (in November 1989). By the time of the third test battery (in November 1990), as a result of teacher attrition and the training of one control teacher in January 1990 (see Annex paragraph 2.1) there were 8 experimental and 7 control teachers left in the study. The number of students, however, was still considered to be acceptable.

Geographical Sample

3.3.17 Prior to his visit Mr Brook thought that it would be preferable to include a wide variety of regions of Nepal within both the control group and the experimental group.

His visit report offered the following five options for selecting schools to be included in the study on a geographical basis:

a. Coverage of all of the schools or a percentage thereof;

b. Restriction of the study to Kathmandu valley;

c. Stratified sampling to represent geographical and topographical variations;

d. Selection within the three most accessible areas;

e. Selection within schools on the main roads.

3.3.18 Following the first consultancy visit, however, it was recommended that owing to logistical constraints, a single accessible region should be used for the study. After discussion with the KELT, it was decided that as 97% of Nepal is rural a rural region would be more appropriate than the region around Kathmandu. The Pokhara region was selected as it was the most accessible rural region. All schools included in the study were accessible within a day by public transport.

Timing of Tests

3.3.19 Mr Brook recommended that performance should be monitored over the three years (grades 8, 9 and 10) leading up to the SLC.

3.3.20 Following the first visit by the consultants, it was decided that the tests should be administered in January 1989, before the start of the school year, and again in October 1989. A further test was envisaged for 1990, in grade 9.

3.3.21 Following their second visit, the consultants reported that the timing of the first test was to be delayed to February/March. The reason for this was that the schools were actually closed in January because SLC examinations were taking place there. The second set of tests was conducted in November 1989 and the final tests were completed in November 1990.

3.3.22 In May 1989 the consultants reported that during 1991 the results obtained by the two

groups of students in the SLC would be compared and, possibly, further tests administered by the baseline project. Neither of these intentions were met, however, because of the attrition of teachers from the control group (see paragraph 3.4.2.).

Assessment Technique

3.3.23 It was proposed by Mr Brook following his May 1988 visit that the assessment should cover four areas, with associated testing instruments as follows:

SKILL TO BE ASSESSED INSTRUMENT

Grammatical Knowledge 50 point Multiple Choice

General Linguistic Ability Series of Comprehension Tests

Reading Skills Reading Comprehension Test

Attitude to English Questionnaire in Nepali

3.3.24 Mr Brook emphasized that the instruments should be reliable, valid, culturally appropriate and pitched at the right level. He recommended that an ELT expert be engaged to design the tests.

3.3.25 The choice of testing techniques of Part I of the tests was finalised during the first consultancy visit. The detailed design of part II was to be deferred to a later date. Their first visit recommended the following tests:

TEST	DESCRIPTION	TASK TYPE	TASK VOLUME
Part 1A	Comprehension	Selective Gap Filling	120/hour
Part 1B	Dictation	Sentence Writing	40/half hour
Part II	Writing	Controlled Writing	30 minutes
		Cued Writing	40 minutes

3.3.26 During their first visit the consultants administered the gap filling test to teachers being trained under the project. This was done in order to ensure that the teachers could confirm that the tests were valid and fair, and it enabled the consultants to revise the tests to remove questions that the majority of the teachers were unable to answer correctly. The dictation test was also piloted, and this enabled the consultants to determine the appropriate length of pauses, to remove words and phrases that caused difficulty, and to improve the design of the answer papers.

3.3.27 Details of the testing are provided in the March 1990 consultancy report. Part I of the test battery was designed to follow the English Syllabuses for grades 7 and 8 in Nepali Upper Secondary Schools so that each group could have an equal chance of completing the tests.

3.3.28 The gap filling test permits a relatively large number of questions to be posed in a relatively short time (compared with a traditional comprehension test). The gap filling test is also effective in measuring reading comprehension. The March 1990 report indicated that the number of gap filling questions was reduced to 100 items in an hour. This reduction was prompted by the experience with the initial trial of the tests on teachers.

3.3.29 The dictation test was included as a test of listening ability. The test was made difficult, in order to be able to measure continuous improvement of ability when repeating the test three times over a two year period, without students reaching a "ceiling" of high marks. As with the gap filling test, it was considered that there was no bias against the control group because the tests were based on material from the course books.

3.3.30 It was expected that, as tests of general proficiency, both dictation and gap filling tests would demonstrate the improved performance of the experimental group relative to the control group.

3.3.31 The second part of the test battery was simplified, as only one writing test was administered rather than the two tests proposed in the initial report. The test was intended to reveal differences in the writing ability of students. Whilst Part I was expected to measure differences in the efficiency of teachers at what they did already, Part II was intended to show more fundamental differences in student understanding that might emerge as a result of the training. The project staff thought that the experimental group would outperform the control group in its ability to communicate orally and in writing. The expense and difficulty of undertaking objective oral tests ruled that out as an option. The KELT project officer and the consultants thought that a written test was a possible way to demonstrate that students were able to use a language creatively (rather than learning sentences by rote).

3.3.32 In addition to student testing, the baseline study also included teacher observation techniques. The purpose of these techniques was:-

a. to identify criterial characteristics discriminating between the classroom practice of trained and untrained teachers involved in the study, in accordance with project staff accounts of training objectives,

b. to indicate the degree of training take up by the experimental group teachers, and

c. to enable the measurement of relationships between pupils' learning gains and some aspects of teachers' practice.

Training in observation was provided by the consultants to New Era staff during the second consultancy visit, and a manual was prepared.

3.3.33 The methodology of the observation exercises is explained in the March 1990 consultancy report:-

a. Details of the school, observer, teacher, class and lesson were noted.

b. The observer coded three five minute periods at fixed times during the lesson into pupil and teacher speech, English and Nepali language, and the use of questions as opposed to other activity.

c. The observer wrote notes describing the lesson in terms of activities, and completed a checklist to note the presence or absence of specific criterial characteristics established earlier as being likely to discriminate between the two groups. These descriptions were used to cross-reference with other data collected.

d. The observer calculated the proportions of spoken English and Nepali used by the

pupils and the teacher.

3.3.34 Three other types of data were collected for the study:-

a. Self-Report: teachers were asked to describe in a report three recent typical lessons that they had given. These were used as additional information on teacher customary practice;

b. Pupils' Work: the local consultants were asked to obtain samples of work from about 5 pupils in each class. These were used to cross check the observational and self-report data;

c. Teacher Interviews: Structured interviews were held with teachers.

3.4 Problems During Implementation

3.4.1 Problems during implementation occurred following the departure of the KELT officer in 1989. The March 1990 report explains problems with the second battery of tests because, following this departure, there was no contact between the SEPELT project and the BC. The consultants had requested, as early as March, assistance in setting up the November 1989 joint observations with New Era and the attendance of the teachers for two days in Pokhara but, on arrival, they found that the necessary arrangements had not been made.

3.4.2 A second problem during implementation was the fact that during the last year of the study, 1990, six of the control group teachers were trained. One was trained in January 1990 and a further five in June. This left only four teachers in the control group. This was despite the fact that the consultants had agreed with the project staff and the Ministry of Education that this should not happen until at least late 1990 (h in paragraph 3.3.12 refers).

3.4.3 In the third battery of tests in November 1990, therefore, it was not considered to be worthwhile repeating the teacher observations. It was decided, however, that the test data could be undertaken for both the control and experimental groups remaining in the study, including the five control teachers who had been trained in June (with only the one trained in January being taken out of the study).

3.4.4 By November 1990, because of holidays, in no case had the pupils received more than a third of a year's English teaching after the training of their teacher under the project - and in most cases it was less than a quarter of a year. The impact of this on the study was not assessed in detail; for example, by comparing the results of those members of the control group who had been trained with those who had not been trained.

3.4.5 For any further tests however, the control group having been reduced to four, it was now considered to be too small to allow meaningful comparisons to be made with the experimental group of eleven. The wastage of the majority of the control group was a major factor in the decision not to follow up the baseline study in 1991 with either an analysis of the SLC results, or an evaluation of the project.

3.5 Main Findings of the Baseline Study

3.5.1 Detailed results of the baseline study in each of the main reports are described in the

Annex. A discussion of the test data results is contained in the June 1991 report, and of the observational data in the March 1990 report. The following assessments are made:-

a. One can be confident that the results do reflect changes brought about by the training, because the statistical analysis takes account of other variables.

b. Furthermore, the ability of teachers was taken into account. The ability of the control group teachers to teach the course was established. The fact that teachers' test scores had no significant relationship with students' scores provides evidence that this was not a major factor in the results. Although improved teacher-English was an objective of the training, this was factored out of the study in order to focus on the effects of changes in teaching practices.

c. Initial standards of student English were very low. The average initial score for the whole sample was just over 3/40 for dictation, 4/100 for gap filling, and less than one student in three was able to write a single correct sentence in English.

d. Given the low initial standards, it is not surprising that gains after about 100 hours of tuition per year were limited.

e. However, the experimental group improved more in the dictation test than the control group. Over the whole period the experimental group's score improved by 17% in the dictation tests as a percentage of total possible marks, compared with 7% for the control group. (Annex paragraph 2.16).

f. The mean score of those students who had a "high" level of initial ability improved in the second year by 19% of total marks in the dictation test in the experimental group, whilst in the control group mean performance deteriorated by 13 points (Annex paragraph 0).

g. Over the whole period of testing, the "high" ability experimental group improved its mean score by 40% as a share of total possible marks, compared with 1% for the high ability control group (Annex paragraph 0). By the third set of tests, the mean score for the experimental high ability group was 60%, and that for the control group was 22% (Annex Table 7).

h. The movements in the scores of the medium and low ability groups was not statistically significant.(Annex paragraph 2.17).

i. The differences in the gap filling test were less marked than the dictation. By the end of year 1 a small but significant difference between the two groups had appeared (Annex paragraph 0). The difference was slightly wider at the end of the second year, but the difference fell just short of being significant, and for both groups the changes in performance from the earlier tests were not significant (Annex paragraph 2.17).

j. Although the project appears to have had no significant impact on the evidence of the gap filling test performance of the groups as a whole, a different picture emerges with disaggregated data. The "high" ability experimental subgroup improved its mean adjusted score by 15 percentage points, whereas the control subgroup performance declined slightly (Annex paragraph 0).

k. The writing tests revealed almost no evidence of improvement in either group. This reflected the fact that writing was not given prominence in the SLC examination (Annex paragraph 2.12).

l. The teacher observational data did suggest real differences in teaching practice between the two groups (Annex paragraph 0), and the consultants concluded that these must indicate real differences in the classroom experience of students, and arise from the training received by the experimental group. Inter-observer reliability measures lent credence to this conclusion. Five caveats were placed on this conclusion, but in view of the marked differences shown by the data, the evaluators considered that these did not undermine the results:

1. The number of observations was insufficient to ensure a complete picture of teachers' customary practice;

2. The reliability of observation would have been greatly improved by the use of paired observations but these were precluded on cost grounds;

3. Few joint observations with the external consultants were possible;

4. The coding of speech utterances can vary by observer, for example in the case of repetitions or false starts, and where speech is rapid;

5. Experimental group teachers may consciously adopt experimental type teaching practices for observed lessons, thus overstating the true average difference in teaching behaviour; on the other hand, the teacher self-report data also provided evidence of difference in teaching practices between the two groups.

3.6 The Use of Baseline Studies in the Monitoring and Evaluation of Education Projects

3.6.1 A major purpose of this evaluation report is to provide lessons to those considering similar studies for future projects. This section of the report attempts to provide some assessment of the efficacy of this baseline study in the context of the Nepal project, and to relate these findings to the value of such studies in general.

Costs

3.6.2 The first point to note is that costs for the baseline study were relatively high in relation to the project which was to be assessed. The overall cost of the baseline study at over ,42,000 (paragraph 0), amounts to 21% of the project cost (excluding UK training awards) of just under ,203,000 (paragraph 0). This is not to say, however, that the costs could not have been reduced either by simplifying the methodology, or making more use of local consultants (see paragraph 3.6.18), or project staff (see paragraphs 3.6.19 - 3.6.21). Furthermore, had the project been extended as initially proposed, the costs of the baseline study would not have been so high relative to the project costs.

Duration

3.6.3 The baseline study was conducted over a relatively short period (less than two years). As

the consultants pointed out, it is possible that students who benefit from trained teachers over a much longer period would exhibit much more marked improvements in performance. In addition it would be valuable to know whether the teaching had any impact on students in their SLC and their future careers.

3.6.4 For a number of reasons, it is preferable that baseline studies should be set up at the same time as the project so that its design can be incorporated in the project design. This provides the opportunity to collect better data to compare pre- and post- project performance of students and teachers, enables data collection to be built into project design including the job descriptions of project staff, and facilitates the identification of control group teachers.

Tests

3.6.5 This particular baseline study attempted to use three basic tests of student performance. The dictation test provided the strongest evidence of project impact. The gap filling test provided weak evidence of project impact but the results were not statistically significant.

3.6.6 The writing test, which was included on the advice of project staff, did not produce valid results because neither group was able to score significantly in the test. The test was abandoned after the second time. The lack of evidence of improvements in writing skills was attributed to the low emphasis placed on writing in the school curriculum. Of the three tests included in the baseline study only one proved able to provide conclusive results, so it is just as well that a range of tests had been included in the study. This experience demonstrates the value of an appraisal prior to the selection of tests, as they must be appropriate to the specific teaching practices and conditions.

3.6.7 The choice of tests should reflect the need to provide a general assessment of language competence but should also include tests of skills which the teacher training programme could reasonably expect to influence. It is not easy, however, to predict in advance what these skills will be. In this case, the performance in the dictation test, which is a test of listening ability, was influenced to a greater extent than that in tests of linguistic competence and grammar (the gap filling and writing tests). In general, it is appropriate that both types of test are included.

3.6.8 It is desirable, as was done in this study, to specify a trained teacher profile at the outset, which should include a list of criterial features to differentiate the trained teacher from the untrained. The training is then designed with the objective of achieving the trained teacher profile. This necessitates a feasibility study at the appraisal stage of a project, but it would help to ensure that the training has valid objectives which are appropriate to the local conditions, subject to refinement through feedback on training implementation.

3.6.9 The teacher observational data provided evidence that the training had been taken up by the teachers, and thus that it was at least possible that the changes in the pupils' scores were due to the training. If the student tests revealed that there was no difference in performance arising from the training, the observational data would be needed to see whether the problem was the failure of the training to affect teaching practice, or the failure of the teaching practice to affect pupil performance. The observations are also useful in project monitoring as they reveal the extent to which teacher training is producing the desired changes in teacher practices, so that the results can be fed back into training design.

3.6.10 A more limited study might proceed without observational data, since as long as one can confirm that the training improves student performance, it is not necessary to explain this in terms of teacher behaviour. Observations could then be undertaken subsequently if no training impact is observed. Even without planned observations, however, it is desirable to specify the trained teacher characteristics that the training is intended to produce.

3.6.11 The consultants gathered data on other variables that might have affected the results, such as class size, the school pass rate in examinations, etc. Taking account of these factors necessitated the use of computer data analysis. The impact of adjusting the test data to take account of these external factors was very small and did not affect the conclusions of the study. This is a conclusion reached with hindsight, however, and without such checks the results would have been open to criticism that the results are not reliable.

Size of Study

3.6.12 A major problem during this baseline study was the fact that several of the teachers in the control group were trained during the course of the implementation of the study. As the whole basis of the study was to compare the performance of students of trained teachers in the experimental group, with that of students of untrained teachers in the control group, this undermined the study considerably and meant that there was no point repeating the collection of teacher observation data at the time of the third tests. It also undermined the opportunity of continuing to monitor the performance of the students for a further year up to the SLC.

3.6.13 The reason for this was the pressure from untrained staff to obtain the training before the teacher training project came to an end. Especially for a long study, it may be difficult to maintain the integrity of a control group as untrained teachers may have less commitment to the study, and even if they do not get trained they may be replaced by trained teachers or switched to a lower level of class by head teachers acting independently of the study. Such factors may be beyond the control of those committed to the implementation of the baseline study. A clear commitment from government, however, at the start of the project, and careful monitoring by the donor (e.g. in ODA's case by a KELT or TCO, or if this is not possible by the British Council or the High Commission/Embassy), should help to reduce these risks.

3.6.14 The number of teachers included in the study was reduced through wastage from 12 teachers in each group in the first set of tests, to no more than 8 teachers in a group in the third set of tests which took place only 20 months later. This suggests the need to take account of teacher wastage rates in deciding the number of teachers to be included in such baseline studies. It may be preferable to increase the number of teachers by sampling within a class (say half of a class). Although attrition of student numbers is also a major problem, this may be a less risky strategy overall.

3.6.15 There are, however, logistical constraints. It is desirable that all student tests should be conducted over a short period of time in order to obtain comparable data (as otherwise the amount of teaching will vary between class samples). Thus a larger number of schools necessitates a larger trained team of testers.

Staffing

3.6.16 This study was implemented principally by a mixture of UK consultants and local

consultants. The experience and qualifications of the UK consultants ensured the high quality of the study, but they also made the study relatively expensive. The consultancy input of 165.5 days amounts to around three quarters of a staff year, and the average unit cost of this was over ,200 per day.

3.6.17 The UK consultants considered that their New Era colleagues had performed extremely well. New Era spent 330 days on the study, which is over 1.5 person years. The cost of this was only about ,4,000, which represents a per person fee of only ,12 per day. This input, therefore, was extremely cost-effective despite the need for the local consultants to be trained and supervised by the UK consultants.

3.6.18 Where local consultants are employed, they should be used for as many of the routine and time-consuming tasks as is feasible. In this study, for example, much of the expenditure was on marking of tests in the UK, whereas a more recent study in Guinea made use of local staff for this. The decision on how much use should be made of local consultants, and project staff, will depend on how much importance is attached to the quality and objectivity of the data.

3.6.19 The study could obviously have been cheaper had there been capacity within the country to supervise it, for example through ODA-funded project staff. The project KELT officer was unable to do this as his duties in providing the training left no time for the baseline study. It had been EVD's intention when planning the study to use the planned additional project staff, but it did not prove possible because two appointments were not made (See paragraph 2.5.1. - 2.5.3). The decision to use UK-based consultants was therefore inevitable.

3.6.20 The use of project staff and their counterparts has many advantages. They can collect evaluation data which otherwise might be lost, and they are made to be externally accountable for its presentation. Counterpart staff are obviously cheaper to use than consultants; additionally, as they may benefit from the experience in terms of training and practice in research methodology, sustainability of the project is likely to be enhanced. Although the framework for data collection by these local staff might be flexible, it should nevertheless be systematic, rigorous and relevant to project objectives, and be written into the project framework.

3.6.21 Project staff should write project objectives in terms of the 'criterial characteristics' that enable evaluators to discriminate between those who have received training and those who have not. These characteristics should derive from systematic needs analysis and may also be made more precise in the early phases of the project. Thus, in this case, the low value of training in writing would have been identified, and the training programme improved. The 'criterial characteristics' provide a focus for data collection, and also require project staff to define their objectives in concrete and precise terms.

3.6.22 Even where project staff and their counterparts are used for the primary data collection and analysis, it is beneficial to employ outside consultants to validate the data. This makes best use of external consultant's neutrality and professional skills, whilst also ensuring unambiguously the accountability of project staff.

Reporting

3.6.23 The presentation of the results of the study is spread over several reports. There is much duplication of information in these reports, and yet no single report contains all the data. The type

of data analysis also varies between the reports.

Use of Results

3.6.24 Whilst the study was able to show that the project had some impact on the performance of students, there are relatively few applications for this information in terms of evaluation, other than to confirm that the project was not totally without benefit. It is not possible to put a monetary or economic value on the gains in student performance.

3.6.25 It is, of course, not straightforward to provide any measure of language training in economic terms, and such an analysis would be likely to add to the already high cost of this type of baseline work. Without an assessment of the use of language in the labour market, however, it is impossible to assess the value of an improvement in language. In the extreme, where English language is not used in the labour market, then a technically highly successful project would have no economic benefit. An analysis of the use of English in the labour market should, therefore, wherever practicable precede any substantial ELT project.

3.6.26 The main benefit of this type of study, therefore, will usually lie in its providing the means for assessing intermediate project impacts, for example by comparing the effectiveness of different elements of a training programme. In this particular project the original plan had been to provide two months' teacher training, but this was changed to one month to double the number that could be trained. Had some teachers been offered two months' training it would have been possible to use a baseline study to compare the impact of the amount of teacher training on the performance of the students.

3.6.27 The development of tests under a baseline study may be particularly important in situations where the public examinations do not provide a good measure of language skills. In countries where the content and rationale of existing national language tests are at variance with the intended outcomes of a training programme, unless the two are harmonised, potential positive results of the training programme could be lost. For instance in Nepal teachers did not use their acquired skills to teach writing because the public examinations did not test such skills in students.

3.6.28 The teacher training project examined by this baseline study has not continued (although this was not foreseen at the time that this project was selected). The results of the baseline study have not, therefore, been of use in shaping the project's future direction. The fact that it was also decided not to evaluate the teacher training project means that the main value of the study lies in the lessons for future educational baseline studies.

Appendix 1

STUDY DOCUMENTATION

1. Current ELT Projects in Nepal, Bhutan and Paraguay (Document 1 main report, Document 2 Background Information, Document 3 Appendix). Mr W P Brook. March 1988

2. Report on a preliminary visit to Nepal. Mr W P Brook. May 1988

3. Nepal Baseline Study. Report on Initial visit to Nepal. Dr C J Weir and Mr J R Roberts. December 1988.

4. Nepal Baseline Study. Report on a Second Visit to Nepal, New Era Training Course and SEPELT Staff Briefing. Dr C J Weir January 1989.

5. Nepal ELT Baseline Study. Analysis of Observational Data Part One: March 1989 Data. Mr J R Roberts. September 1989.

6. Nepal Baseline Study Interim Report (Monitoring Visit October 31st - November 15th). December 1989.

7. SEPELT Teacher Training Project Baseline Study. Dr C J Weir and Mr J R Roberts. March 1990.

8. SEPELT Baseline Study Nepal: Final Report. Dr C J Weir and Mr J R Roberts. June 1991.

Appendix 2

CHRONOLOGY OF STUDY

February 1988 EVD engages Mr Brook to set up ELT baseline studies.

May 1988 Preliminary visit of Mr Brook to Nepal.

June 1988 EVD Contracts the Consultants to investigate the feasibility of the study.

Nov 1988 First visit by the Consultants to Nepal. Methodology negotiated with the KELT officer and the local consultants New Era.

January 1989 Second visit by the Consultants. Staff of New Era trained, and future visit to select teachers for baseline study planned.

March 1989 First tests and observations by 4 New Era staff; data delivered to Consultants in UK.

November 1989 Monitoring Visit by Consultants, including joint observations, and administration of second tests.

December 1989 Interim analysis of data from Nepal and report.

March 1990 Report including analysis of all available data from first two sets of tests.

November 1990 Third administration of tests.

January 1990 Data submitted to Consultants.

May 1991 Completion of Marking and Analysis of Test Results.

June 1991 Submission of Final Report by Consultants.

ANNEX: BASELINE STUDY RESULTS

1. Introduction

1.1. The results of the baseline study were presented in the consultants' final report, dated June 1991 (Weir and Roberts), in its earlier reports of March 1990 (Weir and Roberts), and the Analysis of Observational Data Part One: March 1989 (Roberts). The purpose of this Annex is to summarize all the data in a single place.

2. Language Assessments

2.1. The first data on student tests were presented in the Weir/Roberts Interim Report of December 1989. No test data were presented in the Roberts March 1989 report because it was decided that it was better to wait for the second set of data so that the results from candidates who had fallen out of the study could be ignored.

2.2. The December 1989 report was written on the basis of data collected during a monitoring visit. No data had been received on 17 tests and 11 observations that took place after the monitoring visit. The March 1990 report was written on the basis of all the data of the first two sets of tests, and the June 1991 report analyzed the language data for the study as a whole.

Students' Language Tests

December 1989 Report: Partial Analysis of First and Second Sets of Test Data.

2.3. The first data, presented in the December 1989 report comprised student data from 7 schools (5 trained and 2 untrained), and 205 students' scripts. The tests included the gap filling reading test, the dictation test, and the two short controlled writing tests.

2.4. It was found that, at the start of grade 8, students' proficiency in English was virtually zero. The results were as presented in Annex Table 4 showing the gain in percentage points, after 115 hours of tuition.

2.5. No significant results were obtained from the writing tests because it was found that almost none of the students could write a complete sentence without mistake, an ability to do so being the basis of the tests.

March 1990 Report: Full Analysis of First and Second Sets of Test Data.

2.6. A further analysis of the second battery of test data is presented in the March 1990 report. This included data of the full sample of 22 schools. The scores of the students for the three tests was first plotted on a scatter graph, and those with exceptional scores were removed, on the basis that such students were unrepresentative of the population. Those scoring more than 15 in the dictation, 24 on the gap filling, or 8 in the writing task in the first test administration were removed. This left 716 students in the sample (343 experimental and 373 control).

2.7. Statistical analysis was used to assess whether the two groups were equivalent, using a computer statistical package "General Linear Models Procedure". The scores were calculated as if initial performance of both the control and experimental group had been identical. Initial scores of the two groups before training were slightly different but these differences were not large and were taken into account in the statistical analysis.

2.8. After taking account of initial differences between the two groups, scores were then adjusted for those other variables which might account for differences between the groups on which complete data were available. These included hours of tuition, size of class and teacher language level.

2.9. The performance of students was also tested for variations in teachers' language ability, but this was found not to be significant (see Annex paragraph 0). The March 1990 report asserts that size of class, number of hours of tuition, SLC results, both generally and in English, made small but insubstantial contributions to the difference of the test results of the two groups.

2.10. The results for the two groups are analyzed in Annex Table 5. In both cases the scores are as if the initial test result was the same for both groups. The adjusted figures show the score as if both groups also had the same figures for the other variables (of class size, etc.).

2.11. The gap between the two groups in the dictation test in the unadjusted figures is 7 percentage points, and adjusting for the other factors reduces this to 6 points. In terms of difference in score (as a proportion of the score of the control group) this amounts to a 40% better score for the experimental group. It was found that the biggest gain took place in those that had a high initial understanding of English.

2.12. The difference between the two groups in the gap filling score was more marginal, and after adjustment for the other factors represents only a one percentage point advantage for the experimental group. This is a 13% higher score for the experimental group, however, as a percentage of the score of the control group and is statistically significant. The writing score is consistently low for both groups which reflects the low priority given to writing in the teaching syllabus.

June 1991 Report: Analysis of All Three Sets of Test Data

2.13. The third set of test data is given in the final report of June 1991. As previously, the students with high scores were removed from the sample, leaving a total of 283 student scripts: 158 experimental group and 125 control group students. A check was made to see if there was a bias resulting from the attrition of different proportions of students of higher or lower ability from the two groups. This was found not to be the case. It was decided not to readminister the writing test at the third battery because the results had not been significant in earlier tests.

2.14. The statistical analysis of the dictation and gap filling tests was repeated, using the three test battery data sets. To take account of other factors explaining differences in the test results, revised data were collected on the SLC general and English pass rates and the number of hours of English taught, and these were used together with the original data on class size and teachers' language level, to adjust the scores. Again the General Linear Model was used, so that the results of the second and third tests can be interpreted as if the initial test scores were identical for the two groups. For all three sets of test data, only the data of those remaining in the

study at the third test were used.

2.15. The results for the whole group are presented in Annex Table 6. The improvement in the performance of the experimental group in dictation between the second and third test was statistically significant, but that of the control group was not statistically significant. In the case of the gap filling test, both groups improved slightly between the two tests and the gap between the two groups increased from 0.04 percentage points in November 1989 to 3.77 percentage points by November 1990. In neither case, however, was the improvement over time statistically significant. Furthermore, the difference between the means of the two samples were not statistically significant either.

2.16. Over the whole period of tests from February 1989 to November 1990, the improvement of the experimental group in dictation, as a percentage of total possible marks, was 17%, compared with an improvement of 7% for the control group. In the case of the gap filling test over a similar period, the improvement in the experimental group was 6.12%, and that for the control group was 2.43%, the latter being just short of significance at the 95% confidence level.

2.17. In addition to examining the performance of each group as a whole, the final report also examines the relative performance of "high", "medium" and "low" ability groups. The "high" ability group comprised those scoring more than 7.5% in the dictation test and more than 5% in the gap filling. The "low" ability group were those scoring less than 5% in the dictation and less than 2% in the gap filling test. It was found that only the "high" ability group data showed a significant difference between the experimental and control group figures. For the other groups, the experimental group improvement was consistently greater than the control group improvement, but the results were not statistically significant.

2.18. The results for the "high" ability group are as presented in Annex Table 7. Whilst the performance of the experimental group improved considerably over the period of the two tests, the performance of the control group actually fell. As identical tests were administered on both occasions the decline is surprising, but the sample size was small and may have reflected 'test fatigue' amongst students asked to complete a test for a third time.

2.19. As with the data for the group as a whole, the improvement over the period of the three tests was calculated for the "high" ability sub-group. The experimental sub-group improvement in dictation over the period of the tests as a percentage of total marks was 40 percentage points, which was statistically significant. The equivalent control sub-group improved by only 1 point in the dictation, which was not significant. For the gap filling tests, the improvement over the three tests was 14.77% for the "high" ability experimental sub-group, and -0.64% for the control sub-group, again with an insignificant trend in the case of the control group.

Teachers' Language Tests:

November 1989 Data

2.20. The December 1989 report gives the results of teachers' language tests for 18 out of the 24 teachers in the study at that time. These results, together with the results of tests administered by New Era staff to the remaining 6 staff (with the exception of the oral test, which could only be administered by the UK consultants) are also included in the March 1990 report.

2.21. Three types of tests were administered:

a. Interviews, based on a British Council "9 band Oral assessment checklist."

b. Student tests - dictation and gap filling

c. A grammar test, used to assess UK University Entrance Language Proficiency.

The results are presented in Annex Table 8.

2.22. The dictation and gap filling test scores, the tests which most closely reflect ability to teach the syllabus, were high for both groups of teachers. In the March 1990 report it was shown that there was no statistically significant difference between the two groups on the gap filling and dictation tests. The conclusion drawn by the consultants, therefore, was that there was little to chose between the groups of teachers in terms of their capacity to teach the content of the school text books.

2.23. There was a more marked difference between the two groups in the grammar test and the oral test. The experimental group scored about one point (out of nine) higher in the oral test. The range was the same for both groups, and all teachers scored at least four which is considered the minimum necessary to be able to teach at secondary level.

2.24. The difference in performance in the grammar tests was found to be statistically significant, and although the untrained group average was brought down considerably by two particularly poor scores, this did not account wholly for the difference. The consultants believed this was probably mainly explained by the fact that the trained teachers had just attended a 28 day course conducted entirely in English.

2.25. The statistical analysis of student performance took into account the data on teacher's performance, where complete data were available. It was found that the teachers' scores had no statistically significant impact on the scores of their pupils.

3. Interviews

3.1. Structured interviews were undertaken with 18 teachers in November 1989, but these results were not analyzed for the December 1989 report. Information was obtained on several features of the teacher (years of training, place of origin, education and training, other occupations), and on the school and class environment (number of hours taught per week and per year, number of pupils, SLC general and English pass rates). The latter set of data was used in the statistical analysis of the language test data (see Annex paragraph 0), but the teacher data were not used as part of the formal baseline study analysis.

4. Observations

September 1989 Report

4.1. The first set of observational data was collected in March 1989 and received by the consultants in May 1989. In September 1989 Roberts completed an analysis of this first set of observational data.

4.2. Four New Era staff undertook observation of six teachers each. Those requiring full details of the methodology adopted should refer to the September 1989 report. The observers collected

four types of data:-

a. Analysis of interaction of teachers with pupils during three five minute samples per class;

b. A global estimate of the ratio of teacher and pupil speech;

c. Field notes describing the lesson;

d. An 11 item check-list of teacher practice, nine of which indicated training uptake, and two of which indicated a lack of training.

4.3. In addition a self report lesson description was obtained from teachers, plus examples of student work.

Observation Data on Use of Language by Teacher and Pupil

4.4. No statistical inference was undertaken of the teacher observation data, because the size of sample was not large enough. The analysis undertaken nevertheless required statistical expertise. The tests included the following steps:-

a. Ratio scores were calculated from the observational data on the use of English and Nepali by both teachers and pupils for the individual classes in the two groups (experimental and control).

b. A number of checks were made to ensure the "consistency" of the data in order to confirm the compatibility of the different types of data and, by so doing, demonstrating that the observations had been conducted properly. Even when particularly high (and possibly unrepresentative) scores for the use of pupil English recorded in some trained teacher classes were removed from the data, the average of the remaining scores is still significantly greater than the average score of the untrained group.

c. For the various ratio scores from the observations of language use, the twenty four classes observed were ranked. Pairs of ranked positions in various ratios of the use of English were plotted on graphs. These comprised a) teacher's English to teacher's Nepali, b) teacher's English to pupil English, and c) pupil English to all other speech.

4.5. The graphical analysis was undertaken in order to see whether, for particular classes, one ratio was unusually high or low in one test compared with performance in the other. The possibility that "clusters" identified on the graph might constitute a grouping was then examined. The use of language in the trained group was compared with that in the untrained group. On this basis it was possible to divide the sample into groups according to the amount of speech and the language used and whether this was mostly spoken by the teacher or the pupil.

4.6. This process allowed the identification of two distinct groups:

a. 8 untrained teachers who clearly encouraged less use of pupil English than any in the trained group;

b. 5 trained teachers whose pupils used more English than any of the untrained

group.

4.7. The remaining seven trained teachers and four untrained teachers in the sample were classified into a further three hypothetical groupings -

c. 4 untrained teachers who "overlap" with the trained group on the use of English measures;

d. 4 trained teachers who "overlap" with the untrained teachers on the use of English measures;

e. 3 trained teachers with limited take-up of training.

December 1989 Report

4.8. It was anticipated in the September 1989 report that further investigations of this type would be made in the second round of data analysis, and that the scores of pupils would be correlated with the groupings identified. However, no further analysis of this type was undertaken in the baseline study. Instead, with the data from the second set of observations, a more straightforward analysis was made of the relative amount of pupil and teacher speech and relative use of English and Nepali language. No observations were undertaken at the time of the third set of tests because of the wastage in the control group through the training of some of the teachers (see paragraph 3.4.5.).

4.9. The December 1989 report provided further data on the use of English from 13 observation sessions completed in November. It was found that, in the case of untrained teachers (four observations), the ratio of pupil's English utterances to the total of all other speech utterances was 10%, over the range 5% to 16%. This compared with an average of 16% in the February/March data, from 12 observations.

4.10. In the case of trained teachers, there was one class where no pupil English was used, but this lesson was incomplete. If this is excluded, the ratio of pupil's English utterances to total speech utterances was 33%, over the range 18% to 55%. This compared with an average of 34%, from data collected from 12 observations in the previous observations (February/March).

March 1990 Report

4.11. Similar decisive differences in the use of English also appear in data presented in the March 1990 report, for all the 22 schools remaining in the study at the second test battery in November 1990. The average percentage of pupil English utterances to all other speech for the 11 experimental schools was 31%, over the range 15% to 50%. For the control group the average was 15% over the range 3% to 35%.

4.12. When these figures were adjusted to take account of all other variables (class size, teachers' language ability) the percentages changed to 30% for the experimental group and 16% for the control group, reducing the difference slightly but leaving a significant remaining gap.

Checklist Data

4.13. The September 1989 report stated that the "checklist" data, whereby teacher characteristics were observed, did not significantly discriminate between the experimental and

control group, except in a single category.

4.14. No clear patterns emerged relative to the hypothetical subgroups identified (see Annex paragraph 0). This was explained on the grounds that a characteristic only has to appear once in order to qualify, so the test is insufficiently sensitive. Furthermore, the categorization of behaviour in the checklist groups was thought to be subjective.

4.15. The December 1989 report also presents data on teacher observation using the checklist of teacher characteristics. This is based on observations undertaken in November 1989, which apparently produced results more indicative of an impact of teacher behaviour than the first set of observations undertaken in March 1989. There was clear evidence that untrained teachers tended to score higher on untrained characteristics, and trained teachers to score higher on trained characteristics.

4.16. The December 1989 report was based on the incomplete data (comprising four untrained teachers and nine trained teachers) that the consultants had been able to collect during their visit. The complete data for the 10 teachers in the experimental group and 11 teachers in the control group were analyzed in the March 1990 report. The raw total for the experimental group of trained characteristics was 60 (67%), with an average of 6 each. For the control group the raw total was 24, with an average of 2.2. (24%). Adjustment of these scores for relevant factors again reduced the gap but did not make it insignificant (with adjusted average scores of 5.88 and 2.14 respectively). Annex Table 9 presents the results. Its source is Annex 8.2 of the March 1990 report.

4.17. The March 1990 report also states that the observation scores of those undertaken by New Era and those undertaken by the consultants were checked for consistency. The correlation coefficient was 0.98. Furthermore, the level of agreement on the checklist items was so great that no further statistical analysis was considered necessary.

4.18. The teacher self-reports and samples of student work were used to validate the observation data. The March 1990 report states that checks against the observational data did not produce any evidence of discrepancies, although the analysis was not quantified.

ANNEX TABLES

Table 1. Secondary English Language Teaching Project Costs

Cost Element in , by Financial Year

Cost Type	**87/88**	**88/89**	**89/90**	**90/91**	**91/92**	**Total**
KELT	80,958	63,931	24,056	1,147	8,127	178,219
Office Asst			445	1,530	293	2,268
Equipment				160		160

Vehicle	14,036					14,036
Consultant				7,992	25	8,017
Total	94,994	63,931	24,501	10,829	8,445	**202,700**

Source: ODA Management Information System

Table 2. Secondary English Language Project-related TC Training Awards

Year of Study	Course	Number of Awards	Average Cost of Award	Total Cost
1988/89	English Language Teacher Education	4	,4,456	,17,824
1988/89	MA - Applied Linguistics	1	,10,530	,10,530
1989/90	Educational Project Planning	1	,8,586	,8,586
1989/90	Teacher Training for ELT	5	,4,523	,22,615
1990/91	Teacher Training for ELT	5	,7,411	,37,055
1991/92	Practice of ELT	2	,6,987[a]	,13,974
Total		18	,6,144	,110,584

a/Estimated Costs as at February 1992.

Table 3. Costs of Study by Financial Year and Cost Component

Year	Costs ,		Cost Component	Total Amount ,
1988/89	,9,760		Preparation	,4,147

1989/90	,29,679		Main UK Consultancy	,34,445
1990/91	,26		Local Consultancy	,3,987
1991/92	,3,114			
Total	**,42,579**		**Total**	**,42,579**

Table 4. February and November 1989: Student Raw Test Results - Gap Filling and Dictation

Test	Experimental Group			Control Group		
	Feb. Score %	Nov. Score %	Change Absol. %	Feb. Score %	Nov. Score %	Change Absol. %
Gap Filling	4%	14%	+10%	5%	9%	+4%
Dictation	5%	8%	+3%	3%	4%	+1%

Table 5. November 1989: Adjusted Student Test Results - Dictation, Gap Filling and Writing.

Test	**Experimental Group**		**Control Group**	
	Unadjusted	**Adjusted**	**Unadjusted**	**Adjusted**
Dictation	21%	21%	14%	15%
Gap Filling	9.08%	8.54%	7.04%	7.54%
Writing (number of correct sentences)	1.11	0.99	0.70	0.80

Table 6. November 1989 and November 1990: Adjusted Student Test Results - Dictation and Gap Filling.

Test	**November 1989**	**November 1990**

	Experimental Group	Control Group	Experimental Group	Control Group
Dictation	20%	15%	28%	17%
Gap Filling	8.03%	7.99%	11.82%	8.15%

Table 7. November 1989 and November 1990: Adjusted High Ability Student Test Results - Dictation and Gap Filling.

Test	November 1989		November 1990	
	Experimental Group	Control Group	Experimental Group	Control Group
Dictation	41%	35%	60%	22%
Gap Filling	20.86%	16.89%	26.63%	11.22%

Table 8. November 1989: Teacher Equivalence Test Results - Oral, Dictation, Gap Filling and Grammar.

Test	Untrained Group		Trained Group	
	Average	Range	Average	Range
Oral-score 1-9	4.7	4 to 7	5.6	4 to 7
Dictation %	90	75 to 100	88	75 to 98
Gap Filling %	78	55 to 92	84	62 to 98
Grammar %	55	25 to 72	70	60 to 83

Table 9. November 1989: Teaching Practice Checklist Results.

Group	Untrained Characteristics	Trained Characteristics
	Score % (Marks/Total)	Score in % (Marks/Total)
Untrained Teachers	41% (9/22)	24% (24/99)

Trained Teachers	0% (0/20)	61% (60/90)