EVALUATION OF A LANGUAGE TEACHING PROJECT IN SOUTH INDIA

ALAN BERETTA

Ph.D.

University of Edinburgh

1986

I hereby declare that this thesis was composed by myself and that the work involved is my own.

Acknowledgements

I am indebted to Dr. N.S. Prabhu without whose constant co-operation this thesis could not have been undertaken. Thanks are due also to the project teachers who went to considerable trouble to provide me with the data for chapters 5 and 6, and to the Department of Linguistics at the University of Lancaster, the Department of Applied Linguistics at the University of Edinburgh, Dr. David Carroll and Dr. Esther Ramani for providing me with either tapes or transcripts of project lessons, which were used in chapter 7.

My Ph.D. Committee, Dr. Alan Davies, Dr. Clive Criper and Ms. Rosamond Mitchell, were particularly helpful in suggesting directions that my research might take. I am especially grateful for the advice and encouragement of my tutor, Dr. Alan Davies, throughout the whole course of this thesis.

Finally, and above all, I would like to thank my wife, Christel, and son, Davy, for putting up with a part-time husband and father for so long.

ABSTRACT

This thesis reports an evaluation of an innovative language teaching project in South India, known as the Communicational Teaching Project (CTP).

A review of the relevant areas of applied linguistic, psychological and educational literatures (Chapters 1 and 2) suggests an approach to program evaluation in which external validity is accorded priority over internal validity, and practical uses that can be made of the evaluation with reference to future programs take precedence over attempts to contribute to theories of language learning.

Shaped by the review chapters and by a critical analysis of the literature surrounding the project (Chapter 3), 4 data-based chapters (4, 5, 6 and 7) examine the project from different perspectives.

Chapter 4 reports a comparison of the effects of CTP teaching and the prevailing structure-based method. Chapters 5, 6, and 7 retrospectively explore what took place in CTP classrooms, looking at levels of implementation, teachers' concerns, and the treatment of error. The data and the analyses are presented in such a way as to facilitate extrapolation by potential users of the CTP from the present study to other circumstances.

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
COLT	Communicative Orientation of Language Teaching
CTP	Communicational Teaching
FA	Factor Analysis
FIAC	Flanders Interaction Analysis
FL	Foreign Language
FLACCS	Florida Cimate and Control System
FLint	Foreign Language Interaction
LI	Level of Implementation
LoU	Level of Use
NRT	Non-Regular Teacher
OScAR	Observation Schedule and Record
RT	Regular Teacher
SoC	Stages of Concern
SoCQ	Stages of Concern Questionnaire

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 0

INTRODUCTION

0. Introduction

0.1 Background to the Study

This thesis reports an evaluation of a language teaching project in South India. The project is known as the 'Bangalore' project or the 'Communicational Teaching Project (CTP)'. It was set up by Dr. N.S. Prabhu of the British Council, Madras in 1979 and it ran for 5 years until 1984. Dr. Alan Davies of the University of Edinburgh was invited in the Spring of 1983 by Dr. N.S. Prabhu to evaluate the project. He accepted the invitation, specifying that he would engage a higher degree student (under his supervision) for the purpose, and that provision should be made for the student to visit the project for an extended period early in 1984. This arrangement was accepted and I was asked by Dr. Davies to take up the brief.

0.2 The Organisation of the Study

The study reported here comprises a review of the relevant literature, a detailed description of the Bangalore project, and data-based analyses of the project.

Chapter 1 reviews the published literature on language teaching program evaluation since 1963. Problems both in conception and practice of these evaluations are highlighted. It is found that program evaluation in the

field of applied linguistics has not been elaborated with sufficient sophistication to guide such inquiry in relation to the CTP. In view of this, Chapter 2 turns to both the educational and the psychological literature to examine the more complex and thorough attention that program evaluation has received in these disciplines. By the end of chapter 2, evaluation comes to be seen more as a matter of external than internal inference, and this perspective shapes the data collection and analysis of Chapters 4, 5, 6, and 7.

Chapter 3 draws from the literature and documentation put out by the project team to present a detailed description of the development of the Bangalore project and its methodology; it also offers a critical review of the literature that has grown up around the project.

The first of the data-based chapters, Chapter 4, reports a comparative inquiry, in which CTP students are regarded as experimental subjects and students who receive the 'regular' structure-based instruction as control subjects. The results of tests administered to both groups are analysed.

In Chapter 5, the principal data source is a set of long narrative accounts of their experience on the project written by CTP teachers; levels of implementation are assigned to each account and correlated with antecedent variables, such as years of experience, duration of involvement with the project, whether or not

the teachers were the normal teachers in the respective schools, and so on.

Chapter 6 explores, through the use of a questionnaire, CTP teachers' stages of concern during their association with the project, and also checks the influence of certain background variables.

Chapter 7, the last of the data-based chapters, reports the analysis of CTP lesson transcripts from the perspective of CTP teacher treatment of pupil error.

The aim of the study, in keeping with the attitude to evaluation derived from the review chapters, is to provide information about the implementation of the Bangalore project and, as far as possible, the effects of the project, in such a way that interested parties, whether researchers, teachers, or administrators may judge the credibility of the findings and extrapolate to other circumstances. In other words, the major purpose of the CTP evaluation is to point the way to better future implementations of similar programs, rather than to contribute to theoretical knowledge about the ways that language is learnt.

CHAPTER 1

LANGUAGE TEACHING PROGRAM EVALUATION: A REVIEW OF THE LITERATURE

1 Language Teaching Program Evaluation: A Review of the Literature

1.1 Introduction

The principal aim of this chapter is to provide a review of previous evaluation studies in foreign language teaching, so that an evaluation of the Bangalore project may be informed by past experience in the field.

Three major publications relevant to the present review appeared in 1963. Firstly, and perhaps most importantly, Campbell and Stanley (1963) published their seminal treatise introducing the concepts of internal and external validity and setting out research designs in terms of both experimental and quasi-experimental investigation. This paper pulled together the field of educational research methodology, and virtually gave birth to an era of more design-conscious inquiry.

Secondly, Carroll (1963) thoroughly reviewed the literature relating to second language research (including what might now be termed 'evaluation') up until that date. Like Agard and Dunkel (1948) and Dunkel (1948) before him, Carroll found most of the research suffering from inadequate attention to design. Dunkel points out that "the weakness of past experimentation is what renders so much of it worthless when it is examined critically" (1948, p.168). He adds:

> we should look aghast at the report of a chemical experiment, or even a recipe, which

> would run something like this: 'I took a little water and heated it until it was pretty hot; then I put in quite a lot of ... ' Though it would be unfair to say that language experiments are usually reported in quite these terms, the tendency is more in this direction than toward the style of scientific precision. (1948, p.168-169).

In a similar vein, Carroll concludes his survey with some despondency: "research has contributed very little to foreign language teaching methodology" (1963, p.1094). In view of this, the pre-1963 studies would be unhelpful in suggesting an approach for the present evaluation study.

Thirdly, Keating's (1963) large-scale comparative evaluation of competing language teaching methods appeared, initiating a disillusionment with evaluation which was quickly compounded by the more famous Scherer and Wertheimer (1964) and Smith (1970) studies, and from which FL program evaluation is perhaps still struggling to recover.

These three publications make 1963 the natural starting point for this review.

By 'program' is meant any classroom treatment, whether a so-called 'method' such as audiolingualism, or a part of a method such as explanation (or perhaps conscious non-explanation) of a grammatical structure. A definition of evaluation is left deliberately vague at this stage to include any reasonably systematic and disciplined attempt at gathering either quantitative or qualitative empirical information. Only in chapter 2 will there be an attempt to specify more clearly a concept of

evaluation.

It should be stressed that the present review does not claim to be exhaustive of all 'experimentation' in foreign language teaching since 1963. The concept of program evaluation in applied linguistics has only recently become more refined, and in many studies since 1963, it was not necessarily seen as a distinct form of inquiry. In view of this, certain criteria were helpful for selection purposes.

Apart from the requirement that a study focus on some form of classroom treatment, a major criterion for selection is that a study should at least attempt to be reasonably disciplined. Thus, for example, the Rassias (1971) approach to teaching which Johnston (1980) tried out, although doubtless a program of some sort (it includes involving students emotionally by, among other things, throwing eggs at them, telephoning them in the middle of the night, etc.), would not be addressed here because Johnston's inquiry could hardly be called disciplined. He compares a group that has been exposed to what he calls the 'intensive language model' with a group whose characteristics are unspecified (except that they had had 40 weeks of language instruction); each group took completely different tests (one took a homemade test involving cartoons, while the comparison group took the MLA Cooperative Foreign Language Test in German); these tests were administered under substantially different conditions. While Johnston admits

that the comparison is not "in any way conclusive" (1980, p.105), he nevertheless contrives to "feel justified in claiming positive initial results for this experiment" (1980, p.106). Such studies as this are outside our present sphere of interest.

Another criterion is availability, which principally, though not exclusively, refers to publication in a widely accessible form. This immediately rules out the countless studies commissioned by governments and government institutions which have not been made available to the research community at large. For example, the British Council and the Overseas Development Agency carry out a large number of evaluations, few of which reach the journals or edited collections; instead, reports are often written for very restricted audiences. Given the restricted audience of this type of evaluation, it is in any case difficult to know whether or not they are scholarly, disciplined investigations.

Probably most evaluations that have been carried out in our field have been for restricted audiences, or have simply been too cumbersome for normal publication outlets. However, a number have reached the potentially interested ranks of researchers, teachers, and administrators, usually because they have dealt with highly politicised issues like bilingual education or because they have fuelled the seemingly never-ending

'methods' debate (which had already wearied Sweet as long ago as 1899). The bilingual evaluations have been reviewed elsewhere (e.g. Swain and Lapkin 1982; Genesee 1983), and almost all of them have adhered to a very similar set of procedures, so reference to them need only be fleeting. The main focus of this review, then, will be on both the classic method studies as well as many lesser known ones. (This is also appropriate as the Bangalore project is more concerned with method than with political issues).

1.2 Organising Principles

In such commentaries on the FL evaluation literature as are available, studies have usually been arranged according to a distinction between 'large-scale' and 'small-scale' inquiry (Carroll 1963; Freedman 1971; Allwright 1972; Von Elek and Oskarsson 1973; Stern 1983). Intuitively, it might be expected that the features presented in Table 1.1 would be indicative of this distinction:

Table 1.1

Features Associated with Large- and Small-Scale Inquiry

Large-Scale	Small-Scale
A Complex of Variables	Single Variables
Long Duration	Short Duration
Natural Variation in Behaviour	Standardised Behaviour
Large Number of Subjects	Small Number of Subjects
Uncontrolled	Controlled

In practice, it is not easy to assign studies to either of these sets of categories unequivocally. As Table 1.2, shows, the bi-polar distinction serves well enough for general reviews, but for a review that is specifically designed to guide an evaluation study, rather more specific categorisation would be required. The recognition that almost any combination of the above features is possible has implications for the research strategy that is adopted.

In Table 1.2, 51 studies are categorised according to the following dichotomies: Complex / Single (C / S), Long Duration / Short Duration (L / S), Natural Variation in Behaviour / Standardised Behaviour (N / S); in addition, the number of subjects (or groups, where stated) is given (No.); and finally, Control is represented by the presence (R) or absence (NR) of

randomisation or matching procedures, although it is recognised that control is influenced by all of the above factors, too.

Table 1.2

Categorisation of 51 FL Studies

Study	C/S	L/S	N/S	No.	R/NR
Keating 1963	C	L (1 yr)	N	5000	NR
Scherer and Wertheimer 1964	C	L (2 yrs)	S	227	R
McKinnon 1965	S	S (9x15 mins)	S	120	R
Asher 1966 (i)	S	S (30 mins)	S	88	R
Asher 1966 (ii)	S	S (30 mins)	S	36	R
Asher 1966 (iii)	S	S (30 mins)	S	32	R
Asher 1966 (iv)	S	S (30 mins)	S	96	R
Asher 1966 (v)	S	S (30 mins)	S	30	R
Wohl 1967	C	L (3 mths)	S	2 gps	R
Casey 1968	C	Ex post facto	N	50	NR
Lim 1968	S	S (7x15 mins)	S	144	R
Sjoberg and Trope 1968	S	S (1 lesson)	S	90	R
Politzer 1968	S	S (6x20-30 mins)	S	4 gps	NR
Chastain and Woerdehoff 1968	C	L (2 semesters)	S	99	R
Xiem 1969	S	S (8x30 mins)	S	20	R
Tucker, Lambert & Rigault 1969	S	S (8 lessons)	S	4 gps	R
Torrey 1969	S	S (15 hrs)	S	48	R
Smith 1970	C	L (4 yrs)	N	1090	NR
Hawkins 1971	S	S (10 weeks)	N	253	NR

Table 1.2 (continued)

Study	N/C	L/S	N/S	No.	R/NR
Hauptman 1971	C	S (3 weeks)	S	69	?
Mueller 1971	C	L (2 semesters)	S	77	NR
Levin 1972 (i)	S	S (6x30 mins)	S	227	R
Levin 1972 (ii)	S	S (6x30 mins)	S	104	R
Levin 1972 (iii)	S	S (6x30 mins)	S	247	R
Levin 1972 (iv)	S	S (6x30 mins)	S	98	R
Levin 1972 (v)	S	S (6x30 mins)	S	170	R
Levin 1972 (vi)	S	S (6x30 mins)	S	57	R
Levin 1972 (vii)	S	S (12x30 mins)	S	577	R
Levin 1972 (viii)	S	S (6x30 mins)	S	235	R
Levin 1972 (ix)	S	S (6x30 mins)	S	152	R
Levin 1972 (x)	S	S (10x40 mins)	S	125	R
Asher 1972	C	S (32 hrs)	N	37	NR
Savignon 1972	C	L (1 semester)	N	42	NR
Fink 1972	S	S (3-5 weeks)	S	27	?
Von Elek & Oskarsson 1973	S	S (10x40 mins)	S	125	R
Olsson 1973 (i)	S	S (6x30 mins)	S	18gps	R
Olsson 1973 (ii)	S	S (6x30 mins)	S	24gps	R
Postovsky 1974 (i)	S	S (12 weeks)	S	50	R
Postovsky 1974 (ii)	S	S (12 weeks)	S	48	R
Asher et al (1974)	C	L (1 semester)	N	69	NR
Green 1975	C	L (3 yrs)	N	101	R
Gary 1975	S	L (22 weeks)	S	50	R
Seliger 1975	S	S (65 mins)	S	58	R

Table 1.2 (continued)

Study	C/S	L/S	N/S	No.	R/NR
Freedman 1976	S	S (1 lesson)	S	500	R
Bushman & Madsen 1976	S	S (10 hrs)	S	41	NR
Winitz 1981	S	S (3 days)	S	120	R
Wolfe & Jones 1982	C	S (12 weeks)	N	79	R
Pal 1982	C	S (12x45-60 mins)	S	37	R
Van Baalen 1983	S	L (1 year)	N	80	NR
Wagner & Tilney 1983	S	S (5 weeks)	S	21	R
Thiele & Scheibner-Herzig 1983	S	S (34 lessons)	S	43	NR

It must be stressed that the allocation of studies to categories is subjective. The definitions and descriptions of each category (in sections 1.2.1 to 1.2.5 below) attempts to explain the basis on which assignment was carried out.

Also, it is evident that the categorisation takes only limited detail into account (e.g. as already mentioned, randomisation is only one indication of experimental control). Each study will be treated in greater detail in section 1.3. The purpose of the categorisation is to provide an organisational framework within which issues in evaluation design may be viewed. (The framework does not imply value judgments about the relative merits of different studies).

1.2.1 A Complex of Variables and Single Variables

A complex of variables refers to a 'method', not in the more limited sense proposed by Antony (1963), but as a concept that embraces theory of language and language learning (approach), selection, organisation, definition and specification of content (design) and classroom techniques and practices (procedure): this is the concept of method suggested by Richards and Rodgers (1982), and it is consistent with the broader concept that Stern (1983) finds appropriate. Thus a study comparing audiolingual habit learning with cognitive code-learning (e.g. Mueller 1971) would be characterised as 'complex'.

By contrast, the question of inductive versus

deductive presentation of certain grammatical structures (e.g. Seliger 1975), identifies one element of a method and thus would be considered 'single'.

Such cases as Mueller (1971) and Seliger (1975) are probably relatively uncontroversial. However, it might be observed that the several studies relating to delayed starts in oral production are treated differently. The justification for regarding Asher (1972), Asher et al (1974) and Wolfe and Jones (1982) as 'complex' rather than 'single' rests on the claims of these researchers that Total Physical Response is not to be seen merely as a beginning phase, but as a way of presenting language throughout a course of instruction, in contrast, for example, to Gary (1975). However, another researcher might classify these studies in another way.

For many scholars, it is the 'scope' of the variables under investigation that distinguish 'large-scale' from 'small-scale' (e.g. Freedman 1976, p.12; Hammerly, 1982, p.638). Thus, the 'complex' versus 'single' distinction seems a useful one to make.

1.2.2 Long Duration / Short Duration

How long a study should continue before a difference might reasonably be expected is dependent on the nature of the change the investigation in question wishes to measure. It was arbitrarily decided that the dividing line between 'long' and 'short' duration would be set at

one semester or 45 hours. The major justification for this is that a semester forms a natural break; it is also a reasonably substantial stretch of learning time.

Eisner (1984, p.451) cites two surveys reporting the median treatment time of studies published in the American Educational Research Journal: 45 minutes in 1978, and 72 minutes in 1981. In language teaching, they have sometimes been a little longer, but as can be seen from Table 1.2, very many of them are no longer than a few hours each: Seliger (1975) took 65 minutes, Lim (1968) 1 hour 45 minutes, McKinnon (1965) 2 hours 15 minutes, and Freedman (1976) a single lesson. Such brief studies are what Eisner has called "educational commando raids" (1984, p.451). Get in. Get the results. Get out.

Smith (1970, p.6) argues that as the mastery of language is a longitudinal process, it is appropriate for a study of it to be longitudinal. He cautions that "often, initially dramatic results favouring method A or procedure B prove premature when assessments are made over a long period of time". Snow, too, advised that "most generalisations about school learning need to be built on research using substantial samples of learning time" (1974, p.281), arguing that it would be unfair to find for or against a method after examining the effect it produces only within an exceedingly brief time span.

On the other hand, long-duration studies cannot easily maintain stability over time; stability of treatment and stability of population. Ostensible

differences in teaching methods might be established at the beginning of a program, but then gradually come to resemble each other, i.e. the treatment may change. Rate of attrition generally affects the control of long-duration studies (though not, incidentally, for Green 1975). Chastain and Woerdehoff (1968) started off with 169 subjects but after 2 semesters were reduced to 99; similarly, Scherer and Wertheimer (1964) began with 285 and were left with 49 after 4 semesters; which is to say, the population may change.

Attrition is not only associated with long-duration studies. Seliger (1975) required students to turn up for 2 lessons, a recall test on the day following the lessons, and a further retention test a few weeks later. From an initial sample of 81, only 58 survived.

As this brief discussion has aimed to demonstrate, duration is clearly an important issue, and warrants consideration in the present review.

1.2.3 Natural Variation in Behaviour / Standardised Behaviour

The criterion here is the extent to which normal classroom settings and procedures are manipulated so as to standardise treatments or control for extraneous variables. Since it is a question of degree rather than absolutes, assignment is made on the basis of a close and subjective reading of the reports.

It is considered that the use of tape-recorded lessons to eliminate the teacher variable and to ensure that the treatments were standardised (e.g. McKinnon 1965; Lim 1968; Levin 1972; Von Elek and Oskarsson 1973; Olsson 1973; Seliger 1975; Freedman 1976) constitutes an extreme form of manipulation. By contrast, it is considered that Keating modified very little indeed, if anything.

However, some studies are not so readily placed. Chastain and Woerdehoff's (1968) study, for example, was in most respects 'natural', but was judged to be 'standardised' since, as the authors clearly indicate, instructional sequences and procedures were modified so as to maintain contrasts that would not normally have been present (1968, p.273-274).

The main reason for attending to the 'standardisation' issue is because it is a principal concern for internal and external validity. (The topic will be discussed in greater detail in chapter 2).

1.2.4 Number of Subjects or Groups

The question of how large an adequate experimental sample should be is not a matter for prescription but for an appeal to rules of thumb. Carroll (1969) considers that a minimum of 20 observations for each treatment would be necessary for an experiment to have sufficient power to reject the null hypothesis.

The unit of analysis is often the individual for the

very simple reason that, as Henning points out, "it is usually easier to find 100 students for a study than to find 100 classes" (1987, p.152). If the individual is the unit of analysis, generalisability is considerably constrained, because we are dealing with a small range of schools, settings, teachers, and so on. If the unit of analysis is the class, then 20 observations of each treatment would greatly enhance external validity (which Campbell and Stanley [1963] consider generalisability to be).

Put simply, a study involving more subjects is likely to be more powerful than one involving very few. Thus, there may be doubts about Wagner and Tilney's (1983) study and also about Bushman and Madsen's (1976) inquiry: both studies involved only 7 subjects per treatment.

1.2.5 Presence or Absence of Randomisation or Matching Procedures

Randomisation is widely seen as a prerequisite for internal validity, following Campbell and Stanley's (1963) discussion. It is argued that most of the factors threatening the internal validity of an experiment can be guarded against by having comparison groups constituted randomly.

It is stressed that the studies assigned to 'R' in Table 1.2 are not uniform in respect of randomisation or

matching procedures. Also, some studies which are judged to be 'NR' are possibly more likely to have randomly distributed subjects than some judged 'R'. Wolfe and Jones (1982), for example, flip a coin to determine experimental and control treatments for only 2 intact groups. It seems doubtful that this procedure would lead to a more random distribution than Savignon's (1972, p.19) "naturally assembled collectives". In such cases, allocation to 'R' and 'NR' categories is fraught with difficulty.

In spite of the difficulties, categorisation has been attempted because of the importance of the issue. It should, however, be borne in mind that the attempt to summarise inevitably results in some distortion.

1.2.6 Design Types

The different types of designs implicit in Table 1.2 and the studies associated with them are presented in Table 1.3 below.

Table 1.3

Relationship between 51 studies and 16 design types

No.	Design Type	Study
1	C L N NR	Keating 1963; Smith 1970; Savignon 1972; Asher et al 1974;
2	C L N R	Green 1975;
3	C L S NR	Mueller 1971;
4	C L S R	Scherer and Wertheimer 1964; Wohl 1967; Chastain and Woerdehoff 1968;
5	C S N NR	Asher 1972;
6	C S N R	Wolfe and Jones 1982;
7	C S S NR	Hauptman 1971 (?);
8	C S S R	Pal 1982;
9	S L N NR	Van Baalen 1983;
10	S L N R	/
11	S L S NR	/
12	S L S R	Gary 1975;
13	S S N NR	Hawkins 1971;
14	S S N R	/
15	S S S NR	Politzer 1968; Fink 1972 (?); Bushman and Madsen 1976; Thiele and Scheibner-Herzig 1983;
16	S S S R	McKinnon 1965; Asher 1966; Lim 1968; Sjoberg & Trope 1968; Xiem 1969; Tucker, Lambert & Rigault 1969; Torrey 1969; Levin 1972; Von Elek & Oskarsson 1973;

	Olsson 1973; Postovsky 1974; Seliger 1975; Freedman 1976; Winitz 1981; Wagner and Tilney 1983;

Note: Under 'Design Type', C or S in first position refers to 'complex' or 'single', L or S in second position to 'long-duration' or 'short-duration', N or S in third position to 'natural variation in behaviour' or 'standardised behaviour', and R or NR in final position to 'randomised' or 'non-randomised'.

If Table 1.3 is compared with Table 1.1, it is clear that in extremes there is a parallel. Design types 1 (CLNNR) and 16 (SSSR) represent modal large-scale and small-scale studies, and a number of studies correspond to these two types, except in relation to the number of subjects - Allwright (1972, p.154) observes that number appears not to be particularly relevant from the point of view of experimental control; this is borne out by the fact that, contrary to the expectations suggested in Table 1.1, Freedman's (1976) tightly-controlled laboratory study involved 500 subjects, while Scherer and Wertheimer's (1964) less rigorous field-study could only count 227 subjects at the end of its first year.

While there may be clear similarities at extremes, it is equally obvious from Table 1.3 that virtually any combination of categories is manifested. This helps to highlight some of the choices that must be made in an evaluation design.

1.3 Review of Studies According to Design Type

In this section, the 51 studies listed in Table 1.2 are reviewed according to the 16 design types of Table 1.3.

Insufficient information with regard to randomisation was reported in Hauptman (1971) and Fink (1972) (hence the question marks following mention of these studies in Table 1.3). They will be discussed

within Types 7 and 15, respectively (although they could just as easily be accommodated within Types 8 and 16). Casey (1968) is the only study in Table 1.2 that is not included in Table 1.3. This inquiry took place when the students were no longer studying, but retrospectively correlated teaching method and achievement on certain tests. It resists categorisation in respect to most of the features in Table 1.2. Its inclusion is warranted in this review because of its attempt to specify distinct teaching methods.

1.3.1 Design Type 1 (CLNNR)

There are 4 studies to be considered within design type 1: Keating (1963), Smith (1970), Savignon (1972), and Asher et al (1974).

Keating (1963) investigated the usefulness of the language laboratory in the teaching of French. More than 5,000 students from 21 school districts participated in this much-maligned study. Freedman (1971, p.33) dismisses it; Smith (1970, p.10) warns that "a careful reading of this study raises serious doubts about the validity of the research". Keating himself makes no claim to the contrary.

In fact, Keating makes it perfectly clear that interpretation of his results is hazardous. He acknowledges that "this study cannot be considered an experiment in any proper sense" (1963, p.24). There is no attempt to specify what kinds of treatment the

experimental subjects received; we are not told to what extent use of laboratories varied, or what use, if any, was made of them atall; and we know nothing about what happened in control classrooms. Again, Keating is perfectly candid about this: "absolutely no provision was made for central control of any kind over the independent language instruction programs going on in the various school districts" (1963, p.38).

His candour is worth noting because, as Stern comments, "the Keating report cause a furore" (1983, p.69). In spite of Keating's frequent disclaimers, Forrester (1975, p.11) finds the study "strongly worded". This is hardly true of the report as a whole, but may be so of the conclusions. Certainly, that was the opinion attributed to Lado, as reported in Smith (1970, p.352):

> Lado believed people are going to read two or three pages of conclusions and ... that the style of these conclusions is too absolute and is not justified.

Referring to the Keating report, Essef, as reported in Smith (1970, p.357) commented that "the last paragraph in that report (page 39) qualifies the results. No-one ever seems to read that paragraph". Lado and Essef read the same paragraph and come away with quite different impressions. What that paragraph actually says is:

> while this study does not purport to demonstrate that the language laboratory cannot be used effectively, it does show that in the schools of the Metropolitan School Study Council, a group of schools characterised by competent and well-prepared teachers, better results in important skills areas are being achieved in instructional

> situations which do not use the language laboratory (Keating 1963, p.39).

Perhaps the severe limitations of the inquiry could have been more boldly proclaimed, but Forrester's accusation that "Keating found the results he wanted" (1975, p.18) is irrational.

Keating tested subjects on listening, reading and speaking skills, using the standardised Cooperative French test for reading and listening comprehension, and a specially prepared Speech Production test. Since there is no information about the instruction received in the different classrooms, it is impossible to guage how far these tests reflected course materials or were program-fair (see Green 1975, p.73). Furthermore, since precisely the same Speech Production test was used at all 4 levels, it might be guessed that its power of discrimination was weakened; in fact, the results show that the difference evident at level 1 is not evident at the other 3 levels.

From the perspective of an evaluation of the Bangalore project, the Keating report is interesting for a number of reasons. Firstly, the investigation took place in a natural setting where both experimental and control groups were engaged in ongoing programs. This would decrease the likelihood of creating a Hawthorne effect, that is to say, there is less chance that one group would be motivated by the knowledge that it was taking part in an experiment. (As chapter 4, section 4.4.2.2 elaborates, the Hawthorne effect is problematic

with regard to the Bangalore project). In Keating's study, there was simply no intervention.

A second area of interest is that in the Keating report it is possible to see some of the major problems in conducting field-studies; they will crop up again and again in later investigations (e.g. Scherer and Wertheimer 1964; Smith 1970). These problems are, in particular, (i) specifying treatments and establishing that they were implemented (also problematic for the Bangalore evaluation; see chapter 4, section 4.4.2.2), and (ii) selecting tests that are reliable, valid and program-fair (see chapter 4, section 4.3.4.2).

The Pennsylvania project (Smith 1970) is probably the most widely reviewed foreign language experiment of all those accounted for in Table 1.2. In 1969, the October issue of the Modern Language Journal and the December issue of the Foreign Language Annals were devoted entirely to reviews of the project.

The study aimed to compare 3 methods, a traditional method (TLM), and 2 methods with an audiolingual orientation, Functional Skills (FSM) and Functional Skills plus Grammar (FSG). These methods were intended to have the following characteristics (cf. Smith 1970, p.22-26):

TLM - Use of L1 predominant; translation exercises; grammatical analysis before application.

FSM - Use of target language predominant; grammar

description rather than prescription, and also incidental to functional skills; adherence to 'natural' order of skills, i.e. listening before speaking, reading before writing; printed materials preceded by oral presentation. FSG - Almost identical to FSM except that provision is made for the formal presentation of grammar.

Results did not show the expected superiority of the audiolingual methods. In fact, the traditional classes were either equal to or superior to the audiolingual classes on all measures after one year (though the differences had largely disappeared by the end of the second year. In view of these unexpected findings, the study was subjected to minute scrutiny and intense criticism. As with Keating (1963), it was perceived that in practice, distinctions between methods were inadequately monitored and that the criterion measures were not program-fair.

To take the latter point first, the tests used at the end of the first year were the MLA Cooperative tests of 1939-1941, which were especially reprinted for the task. Commenting on this, Levin remarks that

> the description of the tests makes it clear that they have an academic orientation that obviously puts TLM at an advantage. During the second year of instruction the 1939-1941 variants were replaced by modern variants, and the differences between TLM and FSM/FSG vanished (1972, p.57).

Valette (1970, p.367) contends that even the modern variants were biased in favor of the traditional students, largely because the measures demanded a

knowledge of vocabulary which was more emphasised in their course materials than in the audiolingual materials. This point is elaborated by Valette (1971, p.825-826):

> it would appear that the better performance on the reading test is a function of vocabulary load, since the traditional approach taught about 1,400 to 1,500 vocabulary items, whereas the functional skills and the modified audiolingual approaches presented only 500 to 600 vocabulary items.

Valette also points out that the tests were in some cases too difficult for certain levels and therefore would fail to discriminate (1969, p.400), that is, the spread of scores was too limted to reflect accurately differences in proficiency. If a test fails to discriminate, it is likely to contribute to a no-significant-difference finding which may mask substantial differences that in fact exist. Evidently, this is not a fault in the test itself, but in its application.

In short, as Lado concluded, "the measuring instruments were not adequate, therefore there was just not a way to find out the differences" (reported in Smith 1970, p.347).

Even more problematic than the tests, perhaps, was the difficulty of establishing that the methods really were distinct. Clark cautions that

> if ostensibly different teaching methods tend in the course of an experiment to resemble one another in terms of what actually goes on in the classroom, the likelihood of finding significant differences in student performance is accordingly reduced (1969, p.392).

Considering the texts used in the comparison classrooms, Valette concludes that the features the traditional and audiolingual texts have in common are more numerous than those which divide them (1969, p.397). Starr (reported in Smith 1970, p.333) believed that all of the textbooks, even the traditional ones, were more or less audiolingual. Moreover, as Clark (1969, p.392) stresses, there was more oral use of the target language than expected in TLM classes and less than expected in audiolingual classes. The picture that emerges is that TLM instructional procedures overlapped into FSM/FSG teaching, and vice-versa.

Aware of the need to monitor implementation so as to avoid the confounding of treatments, Smith developed and used rating scales (Smith 1970, p.304-305). Of all of the studies listed in Table 1.2, this is the only attempt to observe classroom behaviour in a systematic and objective fashion. However, as Valdman (reported in Smith 1970, p.335), Clark (1969, p.392) and Carroll (1969) have remarked, different scales were used for the different methods, thereby precluding comparison. For example, in the rating scale for adherence to the TLM design, no provision is made to measure the amount of target language used in the classroom, although there is such provision for the FSM/FSG methods.

Finally, and incidentally, it is noted that although in Tables 1.2 and 1.3, the Pennsylvania study is judged 'NR', there was in fact some randomisation: in the first

year, 104 schools were assigned to treatments, but this was dependent on the availability of language-laboratory equipment, so randomisation only occurred between the two functional skills groups. Such a sampling bias may well have interacted with the variables under study (see Otto 1969; and Wiley 1969).

The Pennsylvania project prompted Hocking (1969, p.410) to take the view that field-studies are uninteresting because they are inherently uncontrollable. Freedman (1971) considered that the project was totally lacking in control but that it was "controlled as far as it was possible to do so" (p.36); her view was that Pennsylvania showed that field-experiment had no future as a form of inquiry (p.36).

Savignon (1972) compared 2 'communicative' programs and a modified audiolingual program. 3 intact classes totalling 42 subjects formed the sample. There were no significant differences on a number of pre-experimental variables (verbal intelligence, language aptitude and high school percentile rank), so that although the classes were not randomly constituted, baseline data were gathered to try to establish the comparability of the groups.

No systematic monitoring of the lessons took place over the 18-week experimental period. Thus, although the different methods are described in considerable detail, we know nothing about how they were actually implemented.

Savignon acknowledges that "the findings are, in their strictest empirical interpretation, applicable only to the particular context in which they were obtained" (1972, p.66), but still makes generalisations beyond that context that are not sanctioned by her data (p.66). Her study suffers in a similar way to the Pennsylvania project (even though it is so much smaller) insofar as it fails to document implementation. Therefore, in their strictest empirical interpretation, her findings are not even applicable in the context in which they were obtained, because it is not certain what is being compared.

Another Type 1 study is Asher, Kusudo and De La Torre's (1974) comparison of a group following a semester's training in a Total Physical Response (TPR) method and 3 control groups whose characteristics are not specified, except that they were at different stages of the 'normal' program of Spanish. Asher and his colleagues attempt to establish the comparability of the comparison groups by administering the Modern Language Aptitude Test (MLAT) (Carroll and Sapon 1959), but due to time constraints, only the experimental students took the test (Asher et al 1974, p.27). Therefore, the experimental students may have had greater language aptitude than the controls; they may have had higher levels of verbal intelligence; they may have been more motivated; we simply do not know as no baseline data were gathered.

Another confounding variable is the use of measuring instruments which were referenced directly to the experimental training and thus could hardly be considered program-fair (see Asher et al 1974, p.29). (The inadequacies of the testing in this 1974 study will be elaborated in a discussion of program-fair testing - chapter 2, section 2.2.1).

Although Asher et al are prepared to cite high significance levels and to describe findings as "extraordinary" (1974, p.28) - after 45 hours of instruction, TPR students do better on a story test than controls who had received 200 hours of instruction - the considerable likelihood of test-content bias and the fact that initial equivalence was not established render all findings virtually uninterpretable.

Asher et al also report the effects of 90 hours of training; however, nearly half of the experimental group dropped out after the 45 hours and although the 16 who remained were given the Pimsleur Spanish Proficiency Test at the end of their training (scoring above the 50th percentile), the controls were not given this test, so no comparison is possible.

Type 1 studies as a whole are characterised by a lack of control. Only Savignon (1972) collects adequate baseline data (showing that it is possible). None of the testing arrangements are program-fair, and none of the studies monitors implementation (although Smith's [1970]

abortive attempt indicates that there is no reason for it to be excluded from consideration). Attitudes to Type 1 designs have been dismissive.

1.3.2 Design Type 2 (CLNR)

The only study in this section is the York study (Green 1975). Essentially, the effect of the language laboratory on overall development in language learning was investigated. The interest was not in some ideal use of the laboratory but typical use which, at the inception of the study, was thought to be an hour per week (this was later confirmed by questionnaire). The laboratory group was compared with a group using a tape-recorder in the classroom and with another group using both laboratory and tape-recorder. From their results, Green and Hawkins (1975, p.203) concluded that the laboratory "appeared to be an ineffective, though common, exploitation of costly equipment".

The York study lasted for 3 years. Asher et al believe that a long-term inquiry precludes control: with reference to their own study, they remark "since the independent variable was long-term training with a complex instructional program there was not the experimental control that is possible with a laboratory problem such as eye-lid conditioning" (1974, p.25). Yet the York study did manage to achieve considerable control in spite of its duration.

There was a less pressing need to monitor implementation because methodology was not at issue, merely the influence of different kinds of hardware. (It might be argued, however, that the use made of this hardware required more than a descriptive account). The control achieved relates to the collection of baseline data, the matching of subjects, the manipulation of the teacher variable, and the attempts to diminish possible Hawthorne effects.

101 subjects in one school were formed into 3 groups matched on 6 pre-experimental measures. The constitution of the groups remained stable over the 3 years (there was no attrition whatsoever). It is possible to be fairly confident that the groups were initially equivalent and that any changes that occurred over time were not due to alterations in the sample.

As for the teachers, they were rotated equally between the 3 treatments each year and over 3 years (cf. Green 1972, p.322). It was perhaps fortuitous that there were 3 teachers, 3 classes, 3 terms in a year and 3 years for the project. However, the evenness of these factors was opportunistically exploited so that the teacher variable could be considered reasonably well controlled. Only 'reasonably' because as 2 dissenting voices point out, teachers may still have been <u>inclined</u> toward one program or another (see Ankers 1974; Freedman 1979).

The York study is also notable in that it took steps to attenuate potential Hawthorne effects. Hawkins (1975,

p.50) explains that

> since two sets - the laboratory groups - would be making weekly visits to the university to use the laboratory, we arranged for the third - the tape-recorder group - to pay a weekly visit to the Language Teaching Centre also.

Not only that but the researchers were conscious of trying to make all groups feel equally 'special' (Hawkins 1975, p.50).

The problem of program-fair testing surfaces in this study too. At the end of each term, tests were administered to all groups. For the first 8 terms, the tests were constructed by the teachers, and since the laboratory and the tape-recorder groups were using the same texts, comparisons were possible (though not for the laboratory plus tape-recorder group which used different materials). In the ninth and final term, the Pimsleur German Proficiency test was used (Green 1975, p.84). It is not clear, however, how far this test reflected the course content of the 3 groups, whether it favoured one or the other, or was insensitive to all three.

The York researchers claim a fair level of internal validity but are dubious about external validity. This seems appropriate. It would be extravagant to make generalisations from one setting. Forrester (1975, p.26) voices concern that "while experimenters are often much more tentative in their conclusions and warn against generalising to other circumstances", not everyone heeds the caveats. Indeed, this is stressed by Green and

Hawkins: "unless the researcher himself puts the question and gives a considered answer, others will be all too ready to make sweeping and ill-considered generalisations for him" (1975, p.193).

A major contribution of this Type 2 study is that it shows that a long-duration field-study need not be inherently uncontrollable. However, the implementation issue was not addressed and the program-fair testing issue was not resolved.

1.3.3 Design Type 3 (CLSNR)

The only study that falls into this category is Mueller (1971). Mueller compared an audiolingual method with a cognitive code-learning method. He manipulated the two methods so that they contained different emphases rather than distinct forms of treatment. This would appear to have gained nothing, but to have increased the likelihood of overlapping programs. As he observes, "the courses emphasising cognitive code-learning also relied heavily on audiolingual concepts in some areas" (Mueller 1971, p.121).

Randomisation was inconceivable since the 3 audiolingual (AL) and 2 cognitive (C) courses did not occur simultaneously, but were staggered as follows (cf. Mueller 1971, p.117):

AL 1	Fall 1966	Spring 1967
AL 2	Spring 1967	Fall 1967

AL 3	Fall 1968	Spring 1969
C 1	Spring 1968	Fall 1968
C 2	Fall 1968	Spring 1969

The report does not tell us about the constitution of each group, stating only that "the student body in each of these courses was essentially similar" (1971, p.118).

What little reaction there has been to this study has been muted. Levin calls the comparisons "poorly controlled survey studies and not experiments in a stricter sense" (1972, p.52); and Von Elek and Oskarsson caution that "far-reaching conclusions should not be drawn from this experiment as it seems to have lacked some important controls" (1973, p.61). Given that it is difficult to discern what is being compared in the first place, it is something of an understatement to quibble about 'far-reaching conclusions'.

Mueller's study does not suggest strengths or weaknesses of a design type.

1.3.4 Design Type 4 (CLSR)

In this section, there are 3 studies: Scherer and Wertheimer (1964), Wohl (1967) and Chastain and Woerdehoff (1968).

Scherer and Wertheimer's (1964) Colorado project is one of the most widely cited studies (along with the Pennsylvania project) of all those referred to in this review. It deserves considerable attention so many years

later because of its interest with regard to evaluation design.

Scherer and Wertheimer started out with the intention of imposing rigorous experimental control, but found that plans very quickly went awry when they were put into operation: "unforeseen factors resulted in some differences between the study as actually planned and as actually carried out" (1964, p.16).

They intended to use a matched pairs design but were forced to abandon it and resort to a coarser matched group design (p.16). The problem was in matching pairs on more than 2 variables: "if we matched on general intelligence and sex, we should probably mismatch on language learning aptitude, previous experience with German and age" (p.35). Also, with less than 300 students to start with (and expected attrition) every time one student dropped out, the other member of the pair would have to go as well. The matching plans changed, then, but they still deferred to basic sampling requirements.

The duration of the study was originally expected to be 2 years of distinct treatment programs, but for practical reasons this was only possible for one year, although measurements were taken for two years (when all students had the same instruction). It is the first year that is interest to us here.

Even in this first year, however, when the comparison groups were ostensibly receiving distinct treatments, Carroll observes that

> the two methods are not as sharply differentiated as they should be in an experiment ... it is doubtful that the theoretical bases for the contrasting teaching methods in the experiment were sufficiently well-formed to make for highly contrasting methods of teaching (1965, p.279).

This conclusion is also reached by Hilton (1969) and Freedman (1971).

Some attempt at standardisation of treatments was made. Weekly staff meetings, interviews with individual teachers and class visits "convinced the [project] leaders that a reasonably uniform mode of instruction was being pursued within each of the two groups" (Scherer and Wertheimer 1964, p.25). In addition, lesson plans were worked out for each semester and teachers were expected to follow them (1964, p.31). Furthermore, assurances are given that no 'traditional' teacher made use of pattern practice drills. Clearly, standardisation was taken seriously (which is unsurprising since Wertheimer was a psychologist and would have been familiar with standard texts insisting on this element). However, observation was not systematic and much of the evidence is based on self-report. At the time the project was operating, systematic classroom observation was in its infancy - Flanders' (1960) model had only just appeared and Medley and Mitzel's (1963) attempt to pull the field together and give it direction came out in the project's second year (when distinct forms of instruction had been terminated). Nevertheless, the Colorado study did not

monitor implementation adequately.

In terms of standardisation, the Colorado project was far more impressive than many later studies. For example, the researchers managed to foresee and avoid the difficulties that a differential coverage of vocabulary in the comparison classrooms would raise; since they planned to use identical tests for both groups, "a reasonably mutual vocabulary was thus virtually mandatory" (Scherer and Wertheimer 1964, p.77). We may recall (from section 1.3.1) that Smith (1970) apparently failed to capitalise on this early awareness.

Specifications of the control group treatment is scant. "Time was devoted to the teaching of the sounds of standard German" (1964, p.75), but we do not know how much time; we find that "pronunciation received some attention" (p.75), but not how much or what kind of attention; we find that "students were trained to understand at least portions of the reading material orally" (p.75) but again without specification (emphases added).

As for testing, no suitable standardised tests could be found, so the researchers produced their own tests. This at least has the advantage that the tests can be sensitive to the instruction received by both groups (even though the decision was taken reluctantly). As we will see in chapter 2, section 2.2.2.2, this strategy does not solve the problem of program-fair evaluation.

So far, nothing in this review of the Colorado

project would suggest that it particularly deserved the criticism that it lacked control. Sampling was well carried out; standardisation probably as well as was possible given the state of the art at the time; testing could at least reflect both programs of instruction. In the major areas that field-studies have since fallen short, Scherer and Wertheimer seem to provide an early (if unrefined) model for this kind of inquiry. In this respect, it is surprising that the study tends to be rejected in tandem with the Pennsylvania project (e.g. Stern 1983).

Having said that, control suffered because of a number of accidents that had nothing to do with the evaluation design, but with the susceptibility that field-studies have to extrinsic factors. Agard and Dunkel (1948, p.5) complained that "railroad and coal strikes deprived us of irreplaceable data". Scherer and Wertheimer were assailed by a catalogue of disasters that must have contributed to their forlorn observation that "we cannot help but wish that a whole year had been available for preparation" (1964, p.79).

First of all, before the experiment had begun, a local newspaper printed an article about the project, so that students exercised pressure to join the experimental groups (1964, p.24). The project staff did not give in. Later, students again insisted they should be allowed to join the experimental classes; again, the project staff

stood firm (p.28). However, the effect on motivation and attitude resulting from this can only be guessed at.

Secondly, the construction of a new language laboratory was to be completed before the first semester began on September 19th. The timing is important because for the first 12 weeks of instruction, the audiolingual groups were to receive only oral/aural training and no reading whatsoever. However, the new laboratory was not ready in time. The old laboratory, meanwhile, was fully booked up "but we arranged to keep it open for the experimental students in the evening and on weekends" (p.26). The new laboratory finally became fully operable on December 9th which was the very day the audiolingual students began their reading phase. Scherer and Wertheimer admit that

> the curtailed laboratory work on the part of the experimental students undoubtedly affected the results of the experiment considerably, but no-one can confidently hazard even a guess as to how seriously things were distorted (1964, p.26).

The third major accident was that there was chaos in the administration of tests at the end of the first semester. Large group examinations previously scheduled by other departments meant that testing in the Colorado study could not be simultaneous as planned, but had to be consecutive, in fact over nine days. This prompts another acknowledgement: "There can be little doubt that some students in the later sections received a little information about the tests from students in the earlier

sections" (1964, p.28).

It has already been noted that in spite of the many merits of this study, there are doubts about implementation and testing. This would make interpretation of the findings tentative. The three accidents would make interpretation distinctly hazardous. (No amount of sophisticated statistical technique can change that, a point that is only worth making because Birkmaier and Lange [1967] state that if factor analysis had been used, the conclusions would have been more valid).

Considerable space has been devoted to this critique of the Colorado project. This is because although the researchers fell well short of their original goal to "draw some definite scientific conclusions about the relative merits of the two methods" (Scherer and Wertheimer 1964, p.12), the study is, as Forrester comments, "well worth the attention of anyone proposing to do research on language teaching, because of the care taken in planning and control" (1975, p.5).

Wohl's (1967) study is not widely reviewed or even mentioned (though see Levin 1972, p.48 49; and Von Elek and Oskarsson 1973, p.65).

2 matched groups were involved in this attempt to compare methods of teaching that only differed in the way that grammatical structures were presented. Although the researcher taught both groups, we do not know how far he

was able maintain the distinctions over the 3 months of the study. As Wohl acknowledges, "the results of this study must be considered inconclusive" (1967, p.16).

Somewhat better known is Chastain and Woerdehoff's (1968) 2-semester study comparing audiolingual and cognitive code-learning methods. The report winds up with bald assertion "in conclusion, the results of this study favored the cognitive code-learning theory" (1968, p.279). However, these results cannot be considered findings. We cannot interpret the results with any degree of confidence. This is because implementation was not monitored and the methods may have overlapped, attrition was problematic and the selection of tests was not argued.

Instructional procedures are described in terms of general strategy, textbooks used and orientation of homework assignments. However, methods may have overlapped; there is certainly room for doubt. For the audiolingual students, "in order not to introduce the variable of the language laboratory no classes were held there" (1968, p.275). However, homework <u>did</u> take place in the laboratory: "habit formation was to be practised in the laboratory as homework" (p.275). Not only that, but the cognitve code-learning groups were allowed to visit the laboratory too:

> the language laboratory was not <u>stressed</u> in these classes. The students had the privilege of visiting the language laboratory, but <u>few</u> did so since they were told such work was not part of

> the course and that such visits would not help their grade in the course (p.275, emphases added).

The researchers did not want the laboratory to be introduced as a variable, and yet the experimental classes used it for homework, and although it was not stressed for the control groups and few visited it, it is evident that some did. However, we are not told how many, how often, or what they did there. In short, the use of the language laboratory was a variable, though unacknowledged and unquantified.

169 students were randomly assigned "by a process of odd and even numbers" (p.269) to experimental and control classes. 70 of these students dropped out in the first semester. This attrition rate, we are informed, "was not atypical of the Purdue population" (p.269). Where attrition is a fact of life, experimentation is particularly difficult. It seems to have had a differential bearing on the constitution of the comparison groups in this study. Chastain and Woerdehoff point out that "in the final sample (i.e. 99 subjects) there was a difference in aptitude which, although not significant, could have influenced the results of the study" (1968, p.269).

The tests adopted for the study were the MLA Cooperative tets of 1963. However, we have no way of knowing to what extent these tests were program-fair. As we will see in chapter 2, section 2.2.2.1, the fact that standardised tests may have impressive reliability,

validity and population statistics does not mean that they are appropriate for comparative evaluations.

Doubts about implementation, attrition and testing mean that the results cannot be interpreted confidently, though Chastain and Woerdehoff do not take this view. Levin is baffled by their interpretation of their study: "somewhat astonishingly, the authors interpret the results as clearly favouring the cognitive code-learning theory" (Levin 1972, p.51).

Chastain (1970) continued the investigation insofar as he administered the MLA tests to the 43 students who remained from the original sample of 169 at the end of another 2 semesters in which "these continuing students were distributed in with students who had not been a part of the study during the previous year" (1970, p.257). In other words, no distinct teaching methods were maintained and the sole purpose of the study was to ascertain whether previous exposure to audiolingual or cognitive methods might have a delayed effect. Perhaps unsurprisingly , no significant differences were found.

Since there were no different treatments for experimental and control subjects, this study is not included in Tables 1.2 and 1.3.

1.3.5 Design Type 5 (CSNNR)

The only study in this category is Asher (1972). The only major difference between this study and Asher et al

(1974)(see section 1.3.1) is that the 1972 study is somewhat shorter (32 hours). The problems with the 1974 inquiry are much the same as those for the first field-experiment.

The tests seem clearly biased towards the TPR group. One of the stories used in classroom presentation to the TPR group is reported in full (Asher 1972, p.135). It is entitled 'Mr. Schmidt goes to the office'. Later in the report, we are informed that one of the tests used to compare TPR and control groups is a "story entitled 'Mr. Schmidt goes to the office'" (p.136). In view of this, it is hardly surprising that experimental students outperformed controls.

This test was administered along with a reading test which showed no significant difference with a control group which had received "about the same amount of training" (1972, p.136). When a comparison is made with controls who have received twice as much training, the tests do not include a reading measure, but instead 7 story tests. The justification for this is:

> Since the comparison was not planned as part of the initial study, it was decided to select the most difficult measures of listening skill available, which were the seven stories administered to everyone in the experimental group [i.e. in the lessons] (Asher 1972, p.136).

Test-content bias seems inevitable.

In section 1.3.1 above, in the current section, and in section 1.3.13 below, the inadequacies of Asher's studies are highlighted. This is worthwhile partly

because some researchers cite these studies as if they could be depended on as clear evidence (e.g. Wolfe and Jones 1982, p.274); partly because some scholars use them to back theoretical claims (e.g. Krashen 1982, p.155, 156, 160); partly, that is to say, because they alert us to the dangers inherent in drawing conclusions that the studies do not sanction; but particularly, as the Asher studies demonstrate very clearly the problems of test-content bias.

1.3.6 Design Type 6 (CSNR)

The only experiment falling into this category is Wolfe and Jones (1982), which is an attempt to test the effectiveness of the TPR method at the secondary school level.

Although categorised as a study that involved randomisation, the procedure amounted only to a flip of a coin to determine which of 2 groups was to be experimental and which control. There was no attempt to collect data on pre-experimental differences or to match subjects on any variables atall. Thus differences in achievement may have been due to initial differences in the groups and not to the treatments.

Half of each experimental lesson was the same as the control program (to the extent that they used the same textbook); the other half was TPR. This procedure introduces a multiple treatment interference (see chapter

4, section 4.4.2.2), which further complicates the already difficult situation in which implementation is not systematically monitored. Thus we cannot be sure what is actually being compared.

1.3.7 Design Type 7 (CSSNR)

Hauptman (1971) is the only study in this section, and since it is so elliptically reported it need not detain us for long.

There is no information about sampling or selection procedures if, indeed, any were carried out. Thus, in this comparison of a 'structural' and a 'situational' method, we do not know whether the treatment or initial differences in groups best explain diiferences in attainment.

Secondly, there is no detailed description of the classroom materials and procedures so that it is difficult to know what is at issue.

1.3.8 Design Type 8 (CSSR)

The only study in this section is Pal (1982), which compares the progress of remedial students taught according to cognitive code-learning and audiolingual methods. From this perspective, the study could be regarded as involving a 'complex' of variables. However, it might as easily be characterised as a Type 16 study, since the only differences in procedure are deductive and inductive presentations of the same 6 structures.

37 college students were divided into 2 groups "on the basis of a common diagnostic test" (Pal 1982, p.152). No significant differences were observed on a t-test, but significant differences were registered when Tukey's 2-sample test was applied. Tukey's test is a non-parametric statistical procedure which would be appropriate if one cannot assume a normal distribution. Robson (1973, p.114) recommends the use of t-tests "unless the distribution is obviously non-normal". Guilford and Fruchter subscribe to this view: "when there is any choice ... we should prefer a parametric test" (1978, p.212). Since Pal does not indicate that the distribution was markedly skewed, the use of a non-parametric procedure (especially when the parametric test has failed to secure a significant difference) requires justification.

1.3.9 Design Type 9 (SLNNR)

The only study judged to fall into the Type 9 design is Van Baalen's (1983) study investigating the effect of grammatical instruction with varying degrees of explicitness.

Teachers were sent questionnaires asking them to specify the method of teaching they used. Van Baalen looked for extreme contrasts with regard to degrees of explicitness in the teaching of grammar, but found none (Van Baalen 1983, p.74). He compromised and opted for a relatively implicit, a relatively explicit group, and a

EX BIBL. UNIV. EDINBURGH.

group somewhere between the two (p.74). Thus the treatment specifications are based not on classroom observation, but on compromises made on interpretations of questionnaires appealing to teachers' introspection. This is a fairly tenuous basis for an experiment. (Incidentally, Casey [1968] also identified different teaching approaches solely on questionnaire responses).

Another relevant area in Van Baalen's investigation is the testing instruments he used. He argues that 2 tests - a story-recall test and a 'picture' test - would elicit natural 'acquired' language.

In the story-recall test, students read a Dutch text and immediately afterwards reproduce it in English as far as memory and the aid of pictures allow. "In this way, the number of structures could be controlled at least to some extent, so that the results would allow of a general comparison" (Van Baalen 1983, p.78; emphases added). 'At least to some extent' is a loose basis for an experimental comparison, however, and it is clear that students might avoid the target structures. This raises the question of the appropriacy of using elicitation instruments as tests. It seems doubtful that the story-recall test would achieve high reliability; indeed, none is reported.

The 'picture' test is adapted from the Bilingual Syntax Measure (Burt, Dulay and Hernandez 1973), and is "supposed to create obligatory contexts for the target structures" (Van Baalen 1983, p.78; emphasis added). Even

casual inspection of the pictures suggests that learners could use structures other than the ones 'required'. This is not mere casuistry; obligatory contexts are hard enough to establish in far more tightly structured instruments than picture tests (or other elicitation measures).

One final point is that if the tests are to be measures of "spontaneous language control" (1983, p.76), then it is perplexing that students are 'made aware' in the instructions preceding the story-recall test: "Also keep paying attention to you English (sic)" (1983, p.95).

1.3.10 Design Type 12 (SLSR)

Design Type 11 did not relate to any of the studies listed in Table 1.2. Design Type 12 is represented by Gary (1975) who hypothesised that a delayed start in oral practice in the initial stages of second language learning would result in result in greater progress in both aural comprehension and oral production. The study is long-term (5 months) and randomised (50 students were randomly assigned to experimental and control programs).

Weaknesses of the study were that the teacher variable was uncontrolled (the same teacher took both classes and might have been predisposed to one method or the other), implementation was not documented, and that the reporting of tests and results.

The latter point refers to the fact that the results

are reported without the aid of tables, and as far as can be ascertained from the description, the only significant difference obtained was on the 'daily' tests of listening comprehension (no significant differences were found on the tests administered after 14 and 22 weeks). Gary considers this sufficient for the strong form of her hypothesis. However, we have no information regarding the validity or reliability of the 'daily' tests; and no explanation is offered as to why significant differences on the daily tests failed to show up on the 14- and 22-week tests.

1.3.11 Design Type 13 (SSNNR)

The only study here is Hawkins (1971). He investigated the relative effect of immediate versus delayed presentation of foreign language script on pronunciation.

4 groups were involved but they were not randomly assigned: "no control was exercised by the writer in the assignment of students to class sections" (1971, p.283); nor were any pre-experimental measures taken. Yet no acknowledgement is made that the results could have been influenced by an initial difference in the constitution of the comparison groups.

The question of duration seems quite arbitrary in this study. The pre-reading phase of the experiment lasted 3 weeks which was thought to be a compromise between a 4-week period "leading to boredom" (Hawkins

1971, p.284) and a two-week period allowing the "introduction of so little of the language as to be of dubious value as far as discrimination between the two approaches was concerned" (p.284). It might equally be suggested that 2 weeks could bore students or that 4 weeks is of dubious value.

The conduct of this study is mainly interesting (from the perspective of the Bangalore evaluation) in that the support and cooperation of the teachers was secured by the simple expedient of informing them of the aims and requirements of the experiment (in contrast to Smith [1970] who kept teachers in the dark about allocation to methods). As is evident in chapters 5 and 6, teachers responded wholeheartedly to requests for their help when the purposes of the inquiry were explained.

1.3.12 Design Type 15 (SSSNR)

No study in Table 1.2 fell into design type 14.

4 studies are relevant to design type 15: Politzer (1968), Fink (1972), Bushman and Madsen (1976), and Thiele and Scheibner-Herzig (1983).

Of Politzer's (1968) study, Seliger comments "it is surprising that this study has not gained more recognition among applied linguists although it purports to test one of the basic tenets of modern language teaching method" (1975, p.4).

Politzer addressed the issue of whether explanation should precede a drill, follow the repetition phase of a drill, occur at the end of a drill, or whether it should be introduced atall.

Lessons were pre-recorded and explanations were spliced into the tapes at appropriate junctures. Thus the teacher variable is obviated but external validity is accordingly diminished.

Students were not randomly assigned to groups but language aptitude measures were taken and initial differences later adjusted with an analysis of covariance (a controversial use of this statistical procedure according to Elashoff [1969]). However, at least an attempt was made to compensate for initial differences.

Duration is decidedly short. It may be asked whether six 20 to 30 minute lessons allow enough time for whatever psychological mechanisms are involved in the development of inductive thought processes; and whether the comparison is fair if students are already familiar with deductive processes.

Fink (1972) gives very little information about his research methodology. He does not even tell us how the 3 groups of 9 students were selected. He makes no particular acknowledgements of the limitations of his experiment, contenting himself with a general disclaimer:

> being fully aware of the complexity of factors that were outside the control of the experiment and the necessity to evaluate any findings with the utmost caution ... (1972, p.281).

There is simply insufficient information to permit interpretation.

Bushman and Madsen (1976) compared "1) full suggestopedia, 2) modified suggestopedia and 3) control" (p.34). Each group received 10 lessons according to instructional methods which are very briefly described. The control group treatment consisted of "a modified audiolingual approach" (p.34) which was considered to be normal in contemporary classrooms; modified suggestopedia incorporated all the features of the full treatment except "music, easy chairs, and the living-room environment" (p.34).

The researchers state that "to limit the Hawthorne effect, all subjects were told that they were part of an experiment, but were not given any other information" (p.34). However, subjects who found themselves in a normal classroom environment with 'normal' teaching might be expected to take a different view of the experiment from subjects introduced to classical music, "a living-room atmosphere with carpeted floors and easy chairs" (p.32). Therefore, simply informing students that they are all part of an experiment cannot be thought to have seriously diminished the Hawthorne effect.

Subjects were not randomly assigned to experimental and control groups. 114 students volunteered to take part. Timetabling difficulties reduced that number to 76, and attrition to 41, so that the 6 groups consisted of

only 7 subjects each. Such small non-randomised groups are susceptible to the threat of initial differences, yet no pre-experimental measures were taken.

2 teachers taught all 3 treatments "in order to control for teacher effect" (p.35). A footnote informs us that one of the teachers "with appropriate temperament and excellent teaching skill is Dr. Robert W. Blair who generously instructed three of the teaching groups discussed in this study" (p.37). (The 'appropriate' temperament is described as the "right" one involving the right "philosophical persuasions" [p.36-37] for teaching suggestopedic classes). It is conceivable that the disposition that is required for suggestopedic teaching is less amenable to audiolingual teaching, and that teacher effect was not controlled for.

The description of research procedures comes under the heading of "A controlled experiment" (p.34). The foregoing paragraphs suggest that it is far from that.

The main interest of the Thiele and Scheibner-Herzig (1983) study is its attention to testing. Aware of "the difficulties of devising a valid means of testing the progress made by two classes taught by different methods" (1983, p.280), the researchers made the following provisions within the 'English test': part 3 focused on the teaching objectives of the control group, while part 4 reflected the teaching objectives of the experimental group; part 2 supposedly captured common teaching object-

ives. (As we will see in chapter 2, section 2.2.2.3, this strategy has been tried in educational research but there are problems associated with the reliance on objectives).

1.3.13 Design Type 16 (SSSR)

Type 16 studies are by far the most common in Table 1.3 and, apart from 'number of subjects', largely conform to the 'small-scale' pattern presented in Table 1.1. 31 studies are relevant here: McKinnon (1965), Asher (1966, 5 studies), Lim (1968), Sjoberg and Trope (1968), Xiem (1969), Tucker, Lambert and Rigault (1969), Torrey (1969), Levin (1972, 10 studies), Von Elek and Oskarsson (1973, 2 studies), Olsson (1973, 2 studies), Postovsky (1974, 2 studies), Seliger (1975), Freedman (1976), Winitz (1981), and Wagner and Tilney (1983).

Since these studies are similar in many ways, they will be treated briefly and only for pertinent areas of research methodology.

Firstly, with regard to randomisation, the Type 16 studies are usually more rigorous. Postovsky (1974) and Sjoberg and Trope (1968) used matched-pairs, random groupings were formed by Freedman (1976) Seliger (1975), Xiem (1969), Winitz (1981) and Wagner and Tilney (1983). However, if one looks more closely, doubts about sampling still arise. For example, Wagner and Tilney (1983) have a pool of 21 subjects whom they randomly divide into 3 groups, thus forming groups of only 7. If we use as a

yardstick Carroll's (1969) rule of thumb about 20 observations per treatment being necessary, then 7 would appear to be inadequate. Also, the 21 subjects are made up of 3 language instructors, 9 music graduates, and 9 adult language learners: this make-up would inhibit generalisability.

A number of the studies in this section investigate a major element of audiolingual and cognitive code comparisons, that is, the effectiveness of types of explanation or lack of explanation of grammatical structures (or, from the learner's perspective, the efficacy of deductive or inductive learning). McKinnon (1965), Lim (1968), Sjoberg and Trope (1968), Xiem (1969), Tucker, Lambert and Rigault (1969), Levin (1972), Von Elek and Oskarsson (1973), Olsson (1973), Seliger (1975) and Freedman (1976) all take this element into account. All of these studies are of extremely brief duration - Sjoberg and Trope (1968) and Freedman (1976) allow only one lesson. As with the Politzer (1968) study (section 1.3.12), it might be asked whether such brief investigations can properly address the issues of inductive and deductive learning.

Seliger, after 65 minutes of learning time in his experiment, feels able to conclude that

> the assumption of discovery method adherents that what is learned by the learner through an inductive process is better retained did not seem to be true in this experiment. If those in favor of an inductive approach are coorect, then

> the opposite of what was obtained should have resulted (1975, p.16).

On the other hand, the assumption that Seliger is making is that whatever processes are involved in inductive learning can be fairly tested within the time-scale of 2 lessons, when other research (e.g. Ausubel 1964) suggests that it takes a very long time to discover grammatical rules autonomously and inductively.

Also, out of all these studies, only Sjoberg and Trope (1968) tested for negative transfer, i.e. after a rule was taught to the 'deductive' students, and they were found to be able to use the rule correctly in sentences they had not seen, they were also given a test in which this rule should not have been applied - the students applied it anyway, which raised doubts about the earlier apparent success of the deductive students.

The desire to control the teacher variable is handled differently by different researchers. McKinnon (1965), Lim (1968), Levin (1972), Von Elek and Oskarsson (1973), Olsson (1973), Freedman (1976) and Seliger (1975) get around the problem by replacing teachers with pre-recorded lessons. This procedure removes the studies from typical classrooms and thus renders impossible any generalisation back to the classroom. It precludes any possibility of external validity. One might also consider the Hawthorne effect in this connection. (Wagner and Tilney's (1983) study is particularly suspect in this respect: not only did they have pre-recorded lessons, but

also they attached electrodes to subjects' scalps during the lessons, to test for alpha brainwave production).

Others, like Postovsky (1974), had the same instructors teach experimental and control classes throughout. It is claimed that "in this manner, the 'instructor variable' was completely controlled" (Postovsky 1974, p.233). However, the fact that the same instructor teaches both approaches does not ensure complete control; the instructor may be more favourably disposed to one approach than to another.

4 studies in this section are interesting from the perspective of program-fair testing: Asher (1966 i - v), Postovsky (1974 i and ii), Von Elek and Oskarsson (1973 i and ii) and Levin (1972 i - x).

Asher has already been mentioned in this connection in sections 1.3.1 and 1.3.5. The five studies he reports in the 1966 paper are different from the 1972 and 1974 studies in that the latter are field-trials and of longer duration. The 1966 studies are all very brief (only 30 minutes each), but test-content bias remains problematic.

Asher (1966i) compares one experimental group who respond physically to commands (for example, if the tape commands them to run to the door, they and their instructor do so) and 3 control groups who do not respond physically; instead, one group watches the instructor perform, the second group listens to an English translation, and the third group reads an English

translation of the command.

Retention tests were given to all students 24 hours after the teaching period and again 2 weeks later, and experimental students performed significantly better. However, the test for the experimental group consisted of physically responding to commands, whereas the control groups all had to write down the English translation of the Japanese. It is quite conceivable that having to write translations is far more difficult than having to react physically, and that therefore the tests were biased against the control groups. The tests do not necessarily reflect greater learning on the part of the experimental group, but possibly only that what they were required to do was easier. It may be that control groups could also have carried out the physical commands. An obvious way of checking this would have been to administer both tests to both groups.

Asher (1966 ii) replicated the first experiment with a different language but with no alteration in testing procedure. Asher (1966 iii and iv) replicated the first two experiments but with children rather than adults. By now aware that the tests might be easier for the experimental groups, Asher thinks of 3 possible reasons for this: (i) the translation hypothesis - that the process of translation may disrupt, (ii) the position-cue hypothesis - that physically responding to a first command (e.g. pick up a book) may help predict the second

utterance (it will have something to do with the book), and (iii) the concurrency hypothesis - this suggests that it may be more difficult to listen and write simultaneously than to listen and act simultaneously (cf. Asher 1966, p.83).

Asher (1966 v) set out to remedy these possible testing defects. In this fifth study, both experimental and control groups observed but did not react physically in the training period. In the retention tests, the experimental students responded physically and control students spoke. Asher still refrains from giving both tests to both groups, but in any case, no significant differences were found. Although it could not be clearly stated that the translation, position-cue or concurrency hypotheses were borne out, it was quite evident that control subjects who did not have to write translations were able to perform on a par with experimental students. A conclusion to be drawn from this is that the dramatic results reported in experiments i to iv were a result of test-content bias.

Concerned about test-content bias, Levin acknowledges that his tests may have been slanted in favour of the control groups (1972, p.127); reviewing Levin's work, Freedman (1979, p.191) also highlights this possibility. Von Elek and Oskarsson (1973, p.147, 149) report considerable difficulties in balancing criterion measures to make them program-fair (their problems are elaborated in chapter 2, section 2.2.2.2.

Postovsky (1974) used tests based on the MLA Cooperative tests of 1963 and notes that they may have been too coarse to get at the differences in treatment. He cautions that "due to grossness of measures, the data generated by this investigation can be interpreted only in general terms" (1974, p.237).

1.4 Summary

This chapter organises a review of the literature relating principally to 'method' studies around the issues of long- and short- duration, complete ('complex') and partial ('single') treatments, natural variation in behaviour and standardised behaviour, randomisation, and number of observations per treatment. It is immediately evident that most combinations of these features are manifested in the 51 studies listed in Table 1.2. Each study is subjected to scrutiny and the merits and weaknesses of special interest in arriving at a design for the Bangalore evaluation are highlighted.

It is always easy to find fault with even the most rigorous studies. In fact, rigour can be perceived as a fault. The point of subjecting studies to criticism has not been to show how poor all previous research has been. Far from it. In fact, one of the most widely maligned studies in the applied linguistic literature (Scherer and Wertheimer 1964) is seen as a study could in many ways serve as a model for current inquiry. The point of the

criticisms has been to increase awareness of the problems involved in any form of evaluation and to see how others have attempted to deal with them. The purpose of this is to find guidance in the conduct of the Bangalore evaluation.

Many questions arise from this review that will be pertinent to our inquiry. Can field-studies be controlled? If not, should they be abandoned in favour of 'laboratory' studies? Are there ways of monitoring implementation adequately? Can the Hawthorne effect be minimised? Is program-fair testing possible? What issues are involved in sampling? Is internal or external validity of primary interest? How can the cooperation of teachers be secured? And so on.

Partial answers to all of these questions can be drawn from the review, but a more complete response is possible if the language teaching studies are set in the context of both psychological research methodologies and educational evaluation. Therefore, in the following chapter (2), the FL studies will be considered as part of a larger enterprise. From the broader perspective that this offers, the likelihood of improving our research design is increased.

CHAPTER 2

PROGRAM EVALUATION: A BROADER FRAMEWORK

2 Program evaluation: a broader framework

In chapter 1, it was noted that most of the published evaluation studies in the FL literature have been comparative 'method' inquiries. These have persisted up to the present; (indeed the Bangalore evaluation, as chapter 4 will elaborate, involves a comparative element). It is possible, however, to take a broader perspective, and to consider a wider range of options open to an evaluation. The central issues may be examined in greater depth so that the evaluation designs selected in the current study are well-informed.

It is fair to say that although applied linguists have more recently addressed these issues, it is usually incidental to another topic (e.g. program design). It is noticeable, for instance, that Richards and Rodgers (1986), in the last chapter of their book, discuss program evaluation, but only generally (though they at least take a broader perspective that includes both qualitative and quantitative inquiry, unlike Long [1984] and Richards [1984]). However, although applied linguists have concerned themselves with program evaluation, not one book devoted to the subject has yet appeared. If we compare this to the 73 titles that just one publisher (Sage, in its complete listing for 1985) can point to in mainstream education, then it is clear that <u>educational research</u> should be consulted. It is also clear that the attention to the conduct of research and to questions of

internal and external inference that is evident in their literature suggests that psychological publications would be worth inspection.

In this chapter, then, FL evaluation will be considered alongside both educational and psychological contributions, so that a more differentiated view of evaluation may emerge.

First of all, a brief overview of the field of educational program evaluation will be presented. This will be followed by detailed review and discussion of the issues that have already surfaced (in chapter 1) as major concerns: (i) program-fair testing (section 2.2), (ii) monitoring implementation (section 2.3), (iii) the advantages and disadvantages of 'field' and 'laboratory' studies (section 2.4). The latter issue will especially involve coming to terms with (a) the roles of internal and external validity and (b) the purposes of evaluation. The aim is to arrive at an attitude to evaluation that will shape the evaluation of the Bangalore project.

2.1 A brief overview of educational evaluation

The explosion of interest in program evaluation occurred in the 1960s. Two reasons are generally offered for this. Firstly, in the wake of the launch of Sputnik in 1957, federal funds in the U.S. were poured into curriculum development in science, mathematics and foreign languages and, eventually, to the evaluation of

these programs; this is why government funding was available for the Pennsylvania study (Smith 1970); (see Stern 1983, p.431 for an account of the consequences of Sputnik).

A second reason is that the 'Great Society' reforms of President Johnson in the U.S.A. led to massive compensatory education programs such as Sesame Street, Head Start and Follow Through. For purposes of accountability, evaluation of these programs was required by law (see Wolf 1987). Kerlinger cites a particular politician's demand for pay-off: "we want N.I.E. [National Institute of Education] to show us that we are getting a bang for the bucks we are spending on educational research" (1977, p.8).

One of the consequences for educational researchers was that they had to develop theories and methodologies of evaluation that would meet the responsibilities thrust upon them. The major influence on evaluation thought until this time was Ralph Tyler's (1949) book Basic Principles of Curriculum and Instruction. (There are many overviews of the field of educational evaluation, but for a recent example see Wolf 1987).

2.1.1 Tyler and Behavioural Objectives

Basically, Tyler's approach, which has since had a tremendous influence on evaluation, involved comparing intended outcomes with actual outcomes. First of all, behavioural objectives are specified, then tests are

developed which reflect all of these objectives. This kind of evaluation was used in the frequently mentioned Eight Year Study (Smith and Tyler 1942).

It is worth noting some implications of this approach. To start with, the tests have to be sensitive to the program's aims. Therefore, standardised tests would be inadequate to the task. Secondly, the comparison of intended outcomes with actual outcomes does not necessitate the setting up of experimental and control groups. Thirdly, and somewhat problematically, the process of arriving at behavioural objectives is fraught with potential misinformation.

It is worth pausing briefly to discuss the role of objectives in evaluation and to note the enduring influence that such a pragmatic approach would have on later evaluators.

Cronbach, who participated in the Eight Year Study, is informative on the issue of how objectives were teased out of the 30 schools in the inquiry:

> As matters turned out, no matter what a school's initial list of goals, each of the thirty local discussions ended with agreement on very nearly the same comprehensive set of objectives.
>
> A teacher who came to a meeting prepared to list the <u>topics</u> of her chemistry course - oxidisation, equilibrium, the halogens - was not allowed to stop there. Was she perhaps also concerned with her students' progress in the use and understanding of scientific method? Did her goals stop with proper use of the metric system and with successful reproduction in the laboratory of results described in the textbook? Or would she also want students to keep good records of observations? To find loopholes in

> arguments? To formulate scientific propositions in testable form? Yes, all those, and the end was not yet. The chemistry teacher found herself led to confess concern that students develop socially while in her charge ... (Cronbach and Associates 1980, p.173-174).

The problem with confining evaluation to behavioural goals is that it ignores unexpected outcomes, outcomes that are hard to define, that are remote in time, difficult to measure; it ignores changes of perception between the time that objectives are stated and the time they are tested; it encourages arbitrariness with regard to continuous outcome variables. Some examples are in order.

McIntyre and Mitchell (1983) remark that the Western Isles Bilingual Project in Scotland had as one of its aims "to instil in pupils a sense of their own identity and to validate their physical, social and cultural environments for them" (p.4). Since this was a long-term objective relating to a general social climate in the Western Isles, it could not be tested. Yet it would appear that the project was motivated by such a goal; in an objectives forum, it would be inadmissible evidence.

Another example: the Bangalore project rests upon an incubation hypothesis (see chapter 3); that is, that acquisition of grammar cannot be forced but will take its own time. Since no deadlines are offered, is the goal to be tested at the end of a semester, two, three, at the end of 2 years, beyond the duration of the project? Again, such a goal is not readily set forth in a testable

manner.

With regard to continuous outcome variables, foreign language programs are particularly susceptible. If a program aims to improve listening skills, it would be reasonable to signal approval whenever scores move along the scale in the direction that has been identified as positive. If a Tylerian evaluator were to ask whether the program has achieved its goals, he would be implying that there is a discontinuity of value on the scale, that there is a point of minimal adequacy; he would be asking for an arbitrary level (see Cronbach 1982, p.221). Knowing what to measure is much easier than knowing the level it should attain.

Patton gives an example of a behavioral objective applied to reading skills: student achievement test scores in reading will increase one grade level from the beginning of the first grade to the beginning of the second grade. He comments:

> this statement is not, however, a goal statement. The goal is that children improve their reading. This is a statement of how that goal will be measured and how much improvement will be desired ... Confusing the (1) specification of goals with (2) their measurement and (3) the standard of desirability is a major conceptual problem in many program evaluations (1982, p.103).

Many goals, then, are abstract, broad, long-term, unplanned, subject to changing perceptions and needs, or relate to continuous scales; as such, they resist transformation into behavioural objectives. Anderson, St.Pierre, Proper, and Stebbins (1978) object that if

such ineffable goals are admissible, program developers effectively put themselves beyond the reach of either corroboration or refutation: "any program that wishes to rid itself forever of the discomforts of evaluation need only add to its list of objectives one metaphysical, obscure, or otherwise unmeasurable purpose" (p.163). In the 1960s, suspicion of unmeasurable goals resulted in bumper stickers bearing the legend "STAMP OUT NONBEHAVIORAL OBJECTIVES" (Atkin 1968). Richards (1984) draws attention to the fact that many currently favoured language teaching methods have not yet been put to the empirical test and insists that they should first of all define their goals. Thus, it is quite clear that in spite of the shortcomings of an insistence on defining and testing objectives, many commentators continue to be influenced by Tyler's approach. While pressing for clear statements of objectives might be useful from the point of view of orienting curriculum development, there are sufficient reasons to doubt its usefulness as a recipe for evaluation.

An aspect of the Tyler approach which has been more generally found wanting is that it ignores process. What happens during the course of a program is irrelevant. The emphasis on test outcomes diverted attention from the 'black box' of the treatment that had been received. As section 2.3 will demonstrate, it is now widely thought that implementation needs to be monitored.

2.1.2 The Rise of Standardised Testing

Evaluators of the 1960s inherited the Tyler model of evaluation. They also inherited the standardised test which made great headway in the 1950s. When compensatory education programs were required by law to evaluate annually, it was specified that they should use standardised test data to determine whether or not projects had achieved their objectives.

In 1947 the Educational Testing Service was established. Following this, documents relating to test standards were formulated by committees of the American Psychological Association (APA 1954; 1966). Standardised norm-referenced testing had laid firm foundations.

The problem for the 1960s evaluators was that a range of objectives relating to a particular program might not be reflected by a standardised test that was designed for more general purposes. (In chapter 1, it was seen that a number of FL studies used standardised tests). The problems associated with the use of standardised testing within evaluation will be pursued in section 2.2.

2.1.3 The Development of the Field

Results from the large evaluations of the 1960s were disappointing (Coleman et al 1966, Cicirelli et al 1969, Ball and Bogatz 1970). It became clear that evaluation was not delivering the goods and that the Tylerian style

of inquiry and the Campbell and Stanley (1963) concept of experimental design were inadequate to the demands made of them. In the 1960s, a few major articles showed how perceptions might change. Cronbach (1963) proposed an emphasis on course improvement; Stake (1967) discussed a 'countenance model' of evaluation which stressed descriptive data and the importance of value judgments; Scriven (1967) made the distinction between 'formative' and 'summative' evaluation (formative being a matter of improving ongoing programs, summative, a question of determining the effects of a program that has come to an end).

The evaluation literature then began in earnest. A number of journals appeared: Evaluation News, Educational Evaluation and Policy Analysis, Studies in Evaluation, CEDR Quarterly, Evaluation Review, Evaluation and Program Planning, New Directions for Program Evaluation and Studies in Educational Evaluation.

Also a plethora of so-called models were developed. These will not be treated in depth here as there are many perfectly adequate reviews available (e.g. Jenkins 1976, Nevo 1983, Fraser 1984, and especially Stufflebeam and Webster 1980) but will be summarised.

There has been 'discrepancy evaluation' (Provus 1971) which slightly elaborates the Tyler model, taking into account the gaps between time-tied objectives and actual performance.

Scriven (1972) proposed 'goal-free evaluation', in which the evaluator pays no attention to stated goals but examines what is actually happening, arguing that if the goals are relevant, they will show up in the classroom; the value of a program resides in the extent to which a program's effects are congruent with the perceived needs of the students.

The 'adversary approach' to evaluation, developed by, among others, Owens (1973) and Wolf (1975), is based on advocacy; teams of evaluators argue opposing points of view and attempt to present a powerful case for their 'side'. Problems with this approach include its cost and the disparity in competence between adversary groups (see House, Thurston and Hand 1984).

Eisner's (1977) concept of evaluation is what he calls 'educational connoisseurship'. No quantitative data is collected; instead the evaluator observes the program in operation and writes a rich, narrative report, in which metaphorical language is encouraged. His approach is also known as the 'art criticism' model. Eisner (1979) gives some examples from his postgraduate students' work; here is one excerpt:

> on some enchanted mornings the contracts take a nap and a different kind of feeling fills the air. On one of these mornings Miss Rogers introduced us to her violin. She began to play it as we sat transfixed, floating from the room through our ears. We drifted to a magic land where sounds change into colors, and the colors are fleecy soft (cited in Fraser 1984, p.130).

Apart from the entirely subjective nature of this

approach, its effectiveness is dependent on how well the evaluator can write.

Contrasting with Eisner's approach, Taba's (1966) 'social studies evaluation model' stresses experimental control, systematic variation of treatments, cause and effect relationships, and statistical analysis. It is based on the view that evaluation is simply an application of standard social science methodology.

Another major 'model' is Stake's (1975) 'responsive evaluation'. In this approach, there is no prearranged evaluation design. Fearing that a prespecified design could lead to narrow and rigid outcomes that may not address the needs of the stakeholding audiences, Stake recommends picking up on whatever turns up and allowing the investigation to be shaped by both the known and unfolding concerns of the stakeholders.

Another approach to be briefly described here is the CIPP (Content, Input, Process and Product) of Stufflebeam et al (1971). The main emphasis here is to provide information for decision-makers. The 'process' element concentrates on implementation (systematic observation, interviews, diaries, participant observation, etc), while the 'product' determines whether or not objectives were achieved.

The process element of the CIPP model is similar to the concept of evaluation espoused by Parlett and Hamilton (1978) which is known as 'illuminative' evaluation. The stress here is on multiperspective

description and triangulation. No 'product' is of interest; 'process' is all.

There are numerous other approaches which have their adherents, but the above summaries represent the best-known. Many of these approaches consciously line up on either side of a divide; they are either qualitative (Guba 1981, Parlett and Hamilton 1978, Stake 1975, Eisner 1977) or quantitative (Taba 1966, Campbell and Stanley 1963). More recently, the perception has surfaced that no single type of evaluation can possibly do service for the wide range of programs that evaluators must address, and the wide range of evaluation purposes. A more eclectic philosophy has emerged which is supported by Weiss (1972), Cook and Reichardt (1979), and most authoritatively by Cronbach and Associates (1980) and Cronbach (1982).

The heterogeneity of evaluation needs and approaches is recognised in the Standards for Evaluations of Educational Programs, Projects and Materials (Joint Committee 1981). Widely shared principles for undertaking evaluations were laid down according to 4 attributes of evaluation: utility, feasibility, propriety and accuracy. The utility atandards relate to the duty of an evaluator to find out who are the stakeholding audiences and to provide them with relevant information on time. The feasibility standards require evaluators to ensure that the evaluation design be workable in real world settings.

The propriety standards demand that the evaluator behave ethically and recognise the rights of individuals who might be affected by the evaluation. Finally, the accuracy standards are concerned with the soundness of an evaluation, requiring that information be technically adequate and that conclusions are linked logically to the data.

2.1.4 Summary

The overview that has just been presented cannot do justice to the issues involved. The purpose has been to illustrate the general trends and the fact that there are now a great variety of approaches to evaluation. The concept of evaluation is still in the process of defining itself, but since the sudden expansion of 20 years ago, evaluation has emerged as a distinct area of inquiry, with its own journals and its own standards.

From the behavioural objectives approach of Tyler (1949), and the rise of standardised testing, a range of qualitative, quantitative and eclectic methodologies has given the evaluator of the 1980s a spectrum of forms of inquiry to select from as the nature of the program to be evaluated requires.

Against this backcloth, specific issues will now be considered.

2.2 Program-fair language teaching evaluation

It was seen in chapter 1 that a perennial problem of

comparative inquiries into FL programs is the likelihood of test-content bias. With reference to both the educational and the applied linguistics literature, the question of program-fair testing within evaluation will now be considered.

A fundamental problem of comparative experiments has been the difficulty of finding or devising tests that could be equally fair to both programs under investigation. As Bathory (1977) asserts, "it is difficult, if not impossible, to construct measurement instruments that are equally valid for different programs" (p.110). This has been recognised for a long time; Agard and Dunkel (1948) argued that "a major obstacle to complete decisiveness of comparative findings is the lack of wholly adequate standards of comparison" (p.13). Walker and Schaffarzick (1974) reviewed 26 studies which attempted to compare curricula and concluded, perhaps unsurprisingly, that "innovative students do better when the criterion is well matched to the innovative curriculum, and traditional students do better when the criterion is matched to the traditional curriculum" (p.94).

This section is concerned, then, with how we get information about the effects of different language teaching programs. First, an illustrative example is given of the dangers inherent in disregarding the predicament. This is followed by a consideration of the

attempts of certain educational and language teaching researchers to come to terms with the need for program-fair assessment

2.2.1 Test-Content Bias

The importance of program-fair testing may be highlighted by an example of the potential for misinterpretation when studies lacking program-fair measures are judged uncritically and cited in support of theoretical positions.

Asher (1972) and Asher , Kusudo, and de la Torre (1974) investigated the effect of the Total Physical Response (TPR) method compared with a 'regular' program. In the 1972 report, one of the stories used in classroom training in the TPR group is presented as an example; it is entitled 'Mr. Schmidt goes to the office'. Later in the report, we are informed that one of the criterion measures used to compare experimental (TPR) and control (regular) groups is a listening test involving a "story entitled 'Mr. Schmidt goes to the office'" (p.136). In view of this, it is hardly astonishing that the experimental students dramatically outperformed controls ($p = .0005$). (On a reading test, no significant differences were found).

These results refer to a comparison with a control group that had received "about the same amount of training as the experimental group" (Asher 1972, p.136). However, when a comparison is made with controls with

more than twice as much training, the criterion measures do not include a reading test in addition to the story test, but instead, seven story tests. The following justifications are given for this:

> Since the comparison was not planned as part of the initial study, it was decided to select the most difficult measures of listening skill available, which were the seven stories administered to everyone in the experimental group (p.136).

Once again, regular students were no match for the TPR group ($p = .005$).

While there is no disputing the results themselves, test-content bias is clearly a possible explanation of them. This is acknowledged by Asher et al (1974):

> It may be argued that an artifact of measurement accounts for the striking differences between groups. Since the stories were developed especially for this project, there may have been an unintentional bias in favor of the experimental training (p.29).

However, they try to argue away the bias by claiming that the stories would be vindicated as reasonable measures if second-semester college students performed better than first-semester college students (p.29). Second-semester students did indeed perform better, but this information has, at best, an oblique and fugitive connection with the bias hypothesis. On the other hand, if stories of the type used in the tests formed part of the TPR training and not part of the control program, then the likelihood of test-content bias is, to say the least, considerable. Quite apart from any other matters of interest regarding

the conduct of the studies, this in itself would render the extraordinary results achieved virtually uninterpretable.

This is worth stressing because on the basis of the TPR studies, large claims are made. For instance, Krashen (1982) considers that "Asher has done a thorough job of putting his method to the empirical test" (p.155). With reference to Asher's 1972 study, he observes:

> Asher reported that after only 32 hours of TPR instruction, TPR students outperformed controls, who had had 150 hours of classtime, in a test of listening comprehension, and equalled controls in tests of reading and writing. Asher's students progressed nearly five times faster! (p.155).

(Incidentally, Krashen here gives the impression that in Asher's 1972 study, TPR students equalled controls who had received 150 hours of instruction on a test of reading and writing, which is quite simply not the case). Krashen adds, in regard to a series of TPR experiments, that they are "clear and consistent, and the magnitude of superiority of TPR is quite striking" (p.156), but we have already seen (in chapter 1, section 1.3.13) that these studies were flawed by test-content bias too. He looks to Asher's work to support his contention that "those methods that provide more of the input necessary for acquisition, and that 'put grammar in its place', are superior to older approaches" (Krashen 1982, p.155).

This rather sanguine gloss of the Asher studies indicates that researchers are not always alive to test-content bias (though see Gibbons 1985, p.259 for a more

cautious view). Of course, the lack of program-fair appraisal is by no means peculiar to the TPR experiments, as will become apparent.

2.2.2 Program-Fair Strategies

A large number of foreign language program evaluations have examined no more than 2 or 3 hours' teaching (as was seen in chapter 1). For short-duration, highly-manipulated studies the question of program-fair appraisal is less pressing, since paper-and-pencil tests are only required to assess a few specified features. So it is mainly to the inquiries of somewhat longer duration that we must turn to see how educational and language teaching evaluators have negotiated the issue.

The principal strategies discernible, which are discussed below, are (a) standardised tests, (b) specific tests for each program, (c) common/unique objectives, and (d) appeal to consensus.

2.2.2.1 Standardised tests

The most common strategy has been to opt for a standardised test. There are two reasons for this: first, they have known characteristics (population statistics, reliability coefficients, item discreteness, and so on); second, and most important, it can be claimed that they are impartial inasmuch as the items are not drawn from either of the programs being investigated, but from an

unspecified universe of content. They have been used in foreign language program evaluation by Keating (1963), Chastain and Woerdehoff (1968), Chastain (1970), Smith (1970), Mueller (1971), Savignon (1972), Postovsky (1974) and Green (1975), among others (cf. chapter 1).

A standardised test was administered at the end of the first year of the Pennsylvania project (Smith 1970) - the specially reprinted Cooperative tests of German and French (Educational Testing Service 1939-1941). Commenting on this, Levin (1972) remarks that

> the description of the tests makes it clear that they have an academic orientation that obviously put TLM [traditional] at an advantage. During the second year of instruction the 1939-41 variants were replaced by modern variants, and the differences between TLM and FSM/FSG [audiolingual] vanished (p.57).

Valette (1970) contends that even the modern variants were biased in favour of the traditional students, largely because the measures demanded a knowledge of vocabulary which was more emphasised in their course materials (1,400-1,500 items) than in the audiolingual materials (500-600 items). (See also Valette 1971, p.825-826). What the Pennsylvania experience demonstrates is that standardised tests are not necessarily impartial.

Not only are they not impartial, but the levels of attainment for which standardised tests are intended may not match the levels achieved by students in specific courses. In the Pennsylvania project, the tests sometimes failed to discriminate; that is, the spread of scores from top to bottom was too limited to reflect accurately

differences in proficiency. If a test fails to discriminate, then it is likely to contribute to a no-significant-difference finding which may mask substantial differences that in fact exist. Evidently, this is not a fault in the test itself, but in its application; nevertheless, when this occurs, a standardised test may not so much be biased in favour of one program as equally unfair to both.

Postovsky (1974), examining the effects of delay in oral production at the beginning of second language learning, administered tests based on the MLA Cooperative foreign language tests (Modern Language Association 1963) but acknowledged that they may have been too coarse. If instruments are insensitive to the learning that has taken place in the comparison classrooms, the risk of misinformed judgment is high. In other words, differences may be accepted as significant when in reality they are not or, conversely, real differences may not be recognised as such. Postovsky cautions that "due to grossness of measures, the data generated by this investigation can be interpreted only in general terms" (p.237). This example shows that standardised tests (or tests closely adhering to them) can contain considerable potential for insensitivity to the features of particular programs.

Concern with standardised tests is constantly surfacing in the educational literature. Koehler (1978)

comments that they "do not necessarily reflect what the teachers were trying to teach" (p.4). Rice and Higgins (1982) found that teachers believe that such tests "are not particularly valid evaluation techniques" (p.18) because they are "unrelated to the curriculum" (p.19). Wood's (1982) survey concluded that "overwhelmingly, teachers prefer to rely on their own observation (and teacher made tests)" (p.18). Marston, Deno, and Tindal (1983) voiced dissatisfaction with standardised tests because of the insensitivity "of these devices in measuring what the student is taught" (p.2). Berliner (1975), reflecting on the impediments to research on teacher effectiveness, points out that because of the intended generality of their application, such tests are inappropriate for classroom-based research: "they simply lack content validity at the classroom level" (p.4). He ventured the opinion that "off-the-shelf standardised tests make poor dependent variables for the study of teaching" (p.5).

The massive Follow-Through (compensatory education program) evaluation is particularly interesting in this respect. House, Glass, Maclean and Walker (1978), who evaluated the evaluation, reported that the program developers repeatedly complained that the standardised instruments employed in the evaluation could not adequately assess the outcomes of their programs. Evidently the developers had originally been promised that special instruments would be used to measure the

diverse effects of their programs, but the evaluators had reneged on the deal. Stebbins, St. Pierre, Proper, Anderson and Cerva (1977) admitted that their outcome measures were closer to the goals of some programs than others, and yet "they continued to compare programs in ways they knew were unfair" (House et al 1978, p.145).

While there is no intention here to attack the use of standardised tests for all purposes (see D.R. Green 1983 for a spirited defence), it is argued that they are entirely unsuitable for the evaluation of instructional programs. The educational literature referred to seems clear on this issue, and the experience of language teaching experimentation encourages similar scepticism. Attractive as they are for their convenience and reliability, they have serious limitations for program comparison.

2.2.2.2 Specific Tests for Each Program

Advocates of standardised tests for program appraisal might argue that the alternatives are even more suspect. Dunkel (1948) believes that standardised tests are desirable or else 'anything goes'. In his view, ad hoc achievement tests will result in research in which

> students are reported to have done well on someone's idea of an adequate test ... if standardised tests are so unsatisfactory, what can we say of the local examination for French II, which Professor X dashed off the night before it was given and which has never been criticised by anyone but him and his students? (p.169-170).

Implicitly sympathetic to this view, Scherer and Wertheimer (1964) nevertheless had difficulty finding appropriate measures. A 2-year search for suitable standardised tests for the relevant level of ability yielded nothing that seemed compatible. In any case, seemingly unconcerned about the potential bluntness of such instruments, they recall that "it had been agreed all along that both groups would take the same tests regardless of the inappropriateness of any portion of the battery for either group" (p.27). They remained convinced that standardised tests were ideal for method comparisons and looked forward to the publication of the MLA Cooperative Foreign Language Tests (Modern Language Association 1963) as if they would ameliorate the kind of testing problems that had confronted them (p.111-112). Thus, they were led, by default rather than by design or conviction, to devise their own tests for each program.

The main shortcoming of this strategy is that since there is no common criterion, direct comparisons are precluded (Shoemaker 1972). Also Method A must be so superior to Method B that significant differences that may show up on A's test are not canceled out on B's. Furthermore, there are a number of possible significant outcomes, most of which would be perplexing. Davies (1983, p.18) details in tabular form the range of possible outcomes of a comparison involving two specific methods; Table 2.1 is an adaptation of this for more

general application.

Table 2.1

Possible Outcomes of a Comparison of Two Methods on Program-Specific Tests

Possible Outcome	Achievement test of Method A	Achievement test of Method B
1	A	A
2	A	n.s.
3	n.s.	A
4	A	B
5	n.s.	n.s.
6	B	A
7	B	n.s.
8	n.s.	B
9	B	B

Adapted from Evaluation and the Bangalore/Madras Communicational Teaching Project (p.18) by A. Davies, 1983, unpublished manuscript, University of Edinburgh, Department of Applied Linguistics. Adapted by permission.

Note: A = Method A students perform significantly better than Method B students; B = Method B students perform significantly better than Method A students; n.s. = no statistically significant difference.

Outcomes 1 and 9 would indicate that one method is superior to another because students perform better not only on their own test, but also on a test relating to the competing curriculum. Outcomes 2 and 8 would appear to suggest that one method is still superior to another, but confidence is decreased. Outcome 5 might cause us to wonder what differences, if any, were manifested in the comparison classrooms. Outcomes 3, 4, 6 and 7 would probably give rise to concern about the tests themselves - Von Elek and Oskarsson (1973, p.159), for example, noted that perceived biases and expected directions of differences failed to materialise on subtests supposedly more in tune with their respective curricula. Even with Outcomes 1 and 9 there is room for doubt. The test for Method A might reflect the method faithfully in every respect, but the corresponding test for Method B might sample the curriculum while asking for different kinds of formats than had been covered in training (Brownell 1966). This is what Hanson, Schutz and Bailey (1977) are referring to when they maintain that "program-specific assessment procedures ... typically have proven too sensitive to one particular program's instruction" (p.1).

The Von Elek and Oskarsson (1973) study, which investigated the relative efficiency of implicit (IM) and explicit (EX) methods of presenting grammatical structures, is a good example of the difficulty of juggling the criterion instruments in deference to the

need to provide for program-fair assessment. Part of the test battery comprised a three-part achievement test in addition to an oral test. The researchers observe that it may be argued that "a test like Part A will favor those who have received an audiolingual training, such as the IM group. Part B was to counterbalance this possible bias" (p.147). They add that it is fair to assume that "Part C with its translation and fill-in items favored the EX group. A corresponding oral test - expected to favour the IM group in corresponding degree - was given" (p.149). Interestingly, in the replication study, the oral test was dropped (p.210), which means that by the authors' own reckoning, the battery must then have been weighted in favour of the EX group.

Considering the merits and shortcomings of constructing specific tests for each program, it is clear that this is no panacea, but it does foster a preoccupation with content validity, without which evaluation may be perceived as illegitimate.

2.2.2.3 Common/Unique Objectives

Another way of attenuating testing difficulties has been offered by Shoemaker (1972), adopted in language teaching research by Thiele and Scheibner-Herzig (1983), and elaborated and operationalised in educational research by Hanson et al (1977) and Hanson and Bailey (1983). In this approach, the objectives of each program are defined, and from these definitions items are

generated that are common to both programs and unique to each.

In their study of kindergarten reading-readiness programs, Hanson et al (1977) used item pools referenced to each program as the elements for constructing 'maxi' and 'mini' tests. Maxi tests were designed to test all the outcomes of a program, and mini tests to measure only the major outcomes. Each group received a composite test comprising a maxi form referenced to its own program and a mini test referenced to a comparison program. Shoemaker (1972) points out that if all outcomes are measured, an instrument could run to hundreds of items. In view of this and on the premise that the class is the unit of analysis, a matrix sampling procedure could be followed, dividing the items among the members of a class.

Attractive as common/unique procedures are for a priori validation, there are a number of obstinate questions associated with them. Most crucially, are specifications of objectives feasible or desirable. It is plain that our perceptions of the potential usefulness of the common/unique protocols will be coloured, at least in part, by our convictions regarding behavioural objectives. And as argued in section 2.1.1, the role of objectives in program evaluation is far from clear.

We can hardly avoid the issue even if, like Hawkins (1975), we consider that "in language teaching, the objectives are implied by the content of the 'syllabus'

and the relative emphasis on the skills involved" (p.37). This would bring us back to content validity, but with little attendant comfort, as even content is not always definable in relevant terms. For example, the Bangalore project dispenses with linguistic curricula and all conscious attention to language, instead proceeding through a series of problem-solving tasks. If content validity involves sampling these tasks, it may be that the resulting tests might probe only peripheral areas. After all, our interest is in language and not in problem solving. The point is that if one program has a linguistic curriculum and another does not, there is not likely to be much commonality in either content or measurable objectives. Thus the common/unique strategy may not offer us a way of comparing them fairly.

2.2.2.4 Appeal to Consensus

The final strategy to be considered is exemplified by Frohlich, Spada and Allen (1985), who hope to be in a position to illuminate issues relating to "the current debate on the respective advantages of more communicative approaches" (p.50). The comparison instruments, however, are prespecified, and this prespecification is external to the claims of particular programs but derived from a model of communicative competence. In effect, what is being said is that there is a broad view, a loose-knit consensus in language pedagogy, that currently holds sway and that this is the model against which all methods are

to be judged. The difficulty here is that if there is a consensus now, might it not change?

While this highly plausible strategy can obviously be useful within certain prescribed limits, it does not offer guidance with respect to the testing dilemma with which we are currently concerned.

2.2.3 Discussion

Hawkins (1975) explains that in the York study, the content and goals of the 3 comparison groups were deliberately aligned to avoid what was perceived as the inevitability of testing bias. He concludes that if the instructional emphasis is different for different groups, "any language test that is used to compare the groups is bound to favour one group more than the other" (p.37).

Certainly, since the strategies described above have not resolved the issue satisfactorily, there are grounds for pessimism. On the other hand, research experience has yielded some basic rules of thumb. We know that comparative program evaluation is only as good as the criterion measures used; that standardised tests are inappropriate tools for comparing programs; and that, at the very least, the claims of each specific program must be taken into account in test construction if competing interests are to be represented fairly.

It may be that specific tests for each program can be supplemented by tests that have some claim to be

syllabus-neutral. That is, tests which could be approached from different perspectives by different groups. Although a priori this would not permit us to be sure that no bias exists, it is at least a reasonable extension of the program-specific testing procedure. (This approach will be further explored in chapter 4).

2.3 Monitoring Implementation

Another major problem for the FL evaluations reported in chapter 1 has been the lack of documentation of the independent variable, i.e. implementation of the programs has not been monitored.

In this section, the literature pertaining to implementation will be reviewed. First of all the need for implementation to be monitored will be fully argued; this will be followed by a discussion of the wide variety of approaches to studying implementation.

2.3.1 The need to monitor implementation

One of the earliest calls for the systematic documentation of program implementation is to be found in language-teaching research. Agard and Dunkel (1948) recommended that a detailed record of classroom procedures should be kept for the following reasons:

> Skeptics often claim that classes are different in name but not in activity. Certainly we have observed reading classes in which very little reading was done and oral classes in which no student spoke. The label is scarcely sufficient warrant for classification. Furthermore, if a group shows outstanding achievement, an

> investigation is primarily interested in discovering hypotheses which may explain its success" (p.7).

Agard and Dunkel's (1948) early insight into the possibility that merely fictional differences might be compared is part of the same awareness that prompted the development of classroom observation schemes a decade or so later (Medley and Mitzel 1958; Flanders 1960) and, more recently, the degree-of-implementation studies of the 1970s and 1980s. Until recently, their advice has not been heeded in the language teaching field where, although treatments have sometimes been described in terms of general activities and materials used, or even casually observed, there has been no systematic monitoring (except for Smith's [1970] abortive attempt, reported in chapter 1, section 1.3.1).

The same has often been true of educational research. Even quite recent studies indicate that the measurement of implementation is by no means routine. Shaver (1983) reviewed articles in the American Educational Research Journal for the years 1969-1981 and found that of 22 teaching-method reports where verification of treatments would have been appropriate, only 9 included such information.

As an illustration, the Follow Through evaluation was widely criticised for failing to measure implementation directly. House et al (1978) took Stebbins et al (1977) to task for classifying program models on the basis of stated goals and objectives: "as anyone

familiar with innovative educational programs knows, some distance frequently exists between a project's stated aims and those it actually pursues" (1978, p.136). Also discussing the Follow Through reports, Fullan (1983) argues in much the same vein:

> Whatever the case, without implementation data from each site we cannot assume that the model and its components have been equally well implemented in different sites. Hence differences in outcomes may very well be related to variations in implementation (p.218).

Charters and Jones put the case for the measurement of the independent variable succinctly:

> what is <u>not</u> standard practice in evaluation studies is to describe, let alone measure, how the programs in experimental and control situations actually differ from one another - or even to certify that they do ... evaluation studies may end up appraising non-events with no-one the wiser (1973, p.5).

Many of the no-significant-difference findings that are to be found in the educational literature may in fact have been comparing programs that were really quite similar in practice. Charters and Jones (1973) examined a number of studies of teaching methods and found that there were no perceptible differences in treatment, and therefore, in Leonard and Lowery's words, "it came as no surprise that non-events led to no significant differences in student performance" (1979, p.4). Similarly , Hall and Loucks believe that

> many of the nonsignificant findings reported in evaluation studies might be better explained if more were known about the actual <u>use</u> of the innovation. In addition, information about the degree of implementation of the treatment might

> be helpful in explaining why certain experimental studies fail to show significant differences or show statistically significant relationships in quite unsuspected directions (1977, p.264).

Hall and Loucks (1977, p268-269) cite an illustrative example in which individualised teaching was compared with non-individualised teaching and in which no significant differences were found. It transpired that 20% of the 'individualised' teachers were not using the appropriate treatment, while 37% of the control teachers were individualising their instruction. The non-significant findings were therefore grossly misleading in that they could have been interpreted as an indictment of the program.

Fullan and Pomfret (1977) call the space between the adoption of a treatment and its termination "a 'black box' where innovations entering one side somehow produce the consequences emanating from the other" (p.337). Anderson et al use the same metaphor, urging that evaluators "must learn not to treat programs as black boxes, with money and theories as inputs and achievement changes as outputs" (1978, p.162). In language-classroom research, Long (1980) uses the same terms to describe the same predicament.

What all of these scholars are saying is that if change is to be measured, it is necessary to know what has changed. This is the most fundamental reason why implementation needs to be documented. In addition, it can provide an opportunity to see why certain projected

changes fail become established, i.e. why they are unimplementable. Moreover, it would be possible to investigate the relative merits of full or partial implementation by correlating it with achievement.

The need for implementation to be studied, then, is well recognised (see also Churchman 1979; Leithwood and Montgomery 1980; Wang and Ellett 1982; Wholey 1979). The question is not so much <u>whether</u> we should measure the independent variable but <u>how</u>. There have been a number of answers to this question, of which the most salient will be discussed in the following section.

2.3.2 Strategies for Monitoring Implementation

As Parlett and Hamilton (1978) point out, process may be documented through a range of procedures: questionnaires, interviews, teacher narratives, participant observation, systematic observation, diaries, rating scales and checklists. A glance at the literature suggests that the most common approach has been to use some form of systematic observation, so our attention will be concentrated on that area. However, some of the other strategies have been tried out in a number of settings and are worth considering.

Hall and Loucks (1977) proposed focused interviews as a means of certifying change, with reference to 8 Levels of Use (LoUs). They argue that implementation is not a yes/no phenomenon but a matter of degree and they

identified 8 levels through 20-minute audiorecorded interviews with teachers. The interview "appears to be a casual conversation about what the interviewee is doing in relation to the innovation" (1977, p.265). Listening later to the audiotapes, two raters assign Levels of Use. If they do not agree, a third rater decides. If the third rater does not agree with either of the original raters, then a larger group focuses on the problem.

This procedure appears to have high inter-rater reliability (from .87 to .96), but it should be borne in mind that these figures are derived from the overall level of use which might obscure considerable variation in segments of use. Concurrent validity of .98 is claimed: ethnographers spent a full day with teachers in the classroom in order to give an impressionistic LoU, which was correlated with the interview ratings. (This approach will be further discussed in chapter 5).

Another approach is exemplified by Stallings (1975). She asked 7 sponsors (program developers) to specify the characteristics of their different curricula in use. After surveying several lessons, she selected for experiment only those in which 'key' elements had been implemented. Implementation scores were derived for each critical variable appropriate to a program model, and then an overall implementation score was computed by totalling the scores across variables. Stallings found that each of the 7 programs was significantly different from controls in terms of treatment. However, she also

found sizable differences between the 4 classes for each program at each site, and also between the 5 or 6 sites allocated to each program. In other words, her implementation study showed that while there may have been between-program differences, there was also considerable within-program variation. This raises the question of the extent to which a program should be standardised, which will be examined in section 2.3.3, but first the strategy of systematic observation is reviewed.

2.3.2.1 Systematic Observation

The aim here is to review types of systems of observation and their development in educational research and then to focus on the adaptations and special requirements in the field of language teaching.

An observation system determines both what observers are to record in the classroom and how they are to do so. It is quite distinct from a rating scale in that it provides a record of teaching behaviours rather than an opinion. It differs from ethnographic observation in that it specifies in advance what behaviours are to be considered, whereas the ethnographer would make post hoc judgments of relevance (see Medley, Soar and Coker 1983).

Systematic observation can be used for any research which requires a record of classroom behaviour, e.g. evaluating teachers, materials and methods, or explaining

outcome differences. It has been used most frequently in process-product research and, more recently, in degree-of-implementation studies. A large number of instruments now exist (see Simon and Boyer 1974), but a great many of them are adaptations (see Rosenshine and Furst 1973, p.138).

Typically, 3 types of observation instrument are distinguished: (i) category, (ii) sign, and (iii) multiple-coding. Category systems code all behaviours into appropriate categories. The categories aim to be mutually exclusive and collectively exhaustive, though to achieve inclusivity, most such systems incorporate a ragbag category, such as Flanders' (1960) 'miscellaneous' category for silence and confusion. Frequencies and sequences of events are recorded and analysed. A sign system, by contrast, codes events that occur within a fixed period (perhaps 5 minutes), but ignoring frequency and sequence and recording each event only once. That it occurs is what is of interest, not how often. Typically, sign systems include a large number of items, e.g. the Florida Climate and Control System (Soar, Soar and Ragosta 1971) has 167 items, which contrasts with the Flanders (1960) category system which has only 10. Multiple coding involves at least two systems recording the same sets of behaviors from different perspectives (an example of the use of multiple coding is to be found in Stallings 1977).

If live-coding is envisaged, each of the three types

of instruments has its advantages and disadvantages. Basically, both category and multiple-coding systems involve fewer and coarser categories so that a coder can record the stream of events as they occur. On the other hand, sequence is preserved so that if, for instance, a teacher praises a student, the immediate result of this might be documented. Sign systems lose the information that can be gained from preserving a sequence, but can take into account a far greater range of behaviours and thus can respond to detail.

If it is possible to work from recordings (audio or video) then there is no need to make the choice between category and sign systems. It would then be feasible to use multiple-codings with large numbers of items. The disadvantage of working from recordings is, as Medley (1982, p.1847) remarks, "the loss of information that results from the reduction in the uses available to the coder when he or she codes from tapes rather than live". This would be especially apparent with audio-recordings. Nevertheless, the advantages of having a permanent record of lessons would seem to outweigh the disadvantages primarily because researchers can then successively refine their perceptions.

Any observation involves perception and judgment on the part of the observer. However, it is clear that some kinds of events require far more interpretation and inference than others. For example, 'shows solidarity'

(Bales, in Simon and Boyer 1970, p.32) necessitates greater judgment than 'the pupil addresses a statement or question to the teacher' (Medley et al, in Simon and Boyer 1970, p.13). Where coding a behaviour depends on extrapolating from a series of events, or on a judgment of something that is nebulous in character, an item would be considered 'high-inference'. Where the behaviour to be coded is overtly specifiable, it is regarded as 'low inference'. The distinction is not absolute but a question of degree. As Borich and Klinzing (1984) comment:

> There is no greater fallacy in direct observation than the belief that all behaviours on a 'low-inference' classroom observation instrument represent comparable levels of inference or even that they are all low-inference (p.37).

As in all forms of behavioural measurement, there is the tension between the objectivity of a criterion and its validity. Low-inference measures are thought to be more reliable because subjectivity is reduced, and high-inference measures are thought to be more valid in that they can take their cue from a number of contiguous events, and can attempt to interpret teachers' intentions. Thus, low-inference items may miss the point while high-inference items may distort it.

Soar and Soar (1982) take the view that since the "behaviors that enter a high-inference rating are not even identified and therefore cannot have their internal consistencies checked" (p.607), they should be avoided.

They conclude that "where it is possible to use low-inference items, the weight of the evidence seems clearly to support their use rather than the use of high-inference measures" (1982, p.607). (However, in their own system - the Florida Climate Control System - as Herbert and Attridge [1975, p.11) note, "the observer is consistently asked to infer the intention of the teacher"). A preference for low-inference schedules is also strongly advocated by Medley (1977; 1978; 1982a).

The position is not simply bi-polar, however. Herbert and Attridge (1975, p.10) make the point that low-inference items may also have considerable potential for distortion, because their selectivity and objectivity may violate the continuity and complexity of behaviour. Therefore, they propose that the key in their criterial statement that "instrument items must be as low in the degree of observer inference required "as the complexity of behavior under study will permit" (1975, p.10; emphasis added) lies in the qualifying clause.

The question of reliability of instruments has brought about a good deal of confusion. Estimating the percent of interobserver agreement is the most frequent procedure used to measure reliability. This involves counting the number of items for which at least 2 coders agreed and then working out the percentage (of the total number of items). Interobserver agreement, as Soar and Soar (1982) suggest, is potentially misleading because "both frequently occurring items and infrequently

occurring items are likely to contribute to percent agreement but are likely to do little to discriminate between teachers" (p.608). Rather than percent agreement, Scott's (1955) coefficient (much criticised by Bailey [1975]) and product-moment correlations have been used. However, there seems to be a consensus that the most appropriate statistical procedure is an interclass correlation (Ebel 1951; Cronbach et al 1963; Frick and Semmel 1978; Soar and Soar 1982; Leinhardt 1980; Medley and Mitzel 1963; Bartko 1976).

There are many researchers who judge that interobserver agreement does not adequately guage reliability (Herbert and Attridge 1975, p.14; Soar and Soar 1982, p.608; McGaw et al 1972, p.14-15). In spite of Medley and Mitzel's (1963) early cautions, many scholars have ignored the limitations and continued to report only percent of interobserver agreement (cf. Bellack et al 1966, p.35). More important is the replicability of the agreement, that is, its stability over observation occasions, because, according to McGaw et al (1972), "unless stable estimates of behavior can be obtained, inter-object variability will inevitably be swamped by intra-object variability" (p.16). That is, there may be more variation <u>within</u> teachers than <u>between</u> teachers.

The procedures involved in estimating reliability, following Medley and Mitzel (1963) and McGaw et al (1972) require multiple observers to be fully crossed with

classrooms and also situations. Rowley (1976) has commented that the magnitude of such a study is

> beyond the resources of most researchers ... consequently, it has been common to ignore the question of reliability altogether, or else to report a coefficient of observer agreement, knowing full well its inadequacy for that purpose (p.51).

Leinhardt (1980, p.408) sympathises with researchers who do not wish the evaluation of an educational treatment to turn into an exercise in validating instrumentation. Even so, as both Rowley (1976) and Leinhardt (1980) point out, stability data are necessary. Thus, as Leinhardt (1980, p.410) argues, it is sensible to maximise the stability of the situations (i.e. the materials, the time of day, the grade-level of the pupils, etc.) in order to avoid a full-scale reliability study (though for a rare account of a full-scale study, see Peterson et al [1985]).

Another issue within systematic observation is whether instruments can be considered generic or subject-specific. For example, Rosenshine and Furst state that "although several systems have been developed for specific subject areas, almost all of these systems can be used in other subject areas" (1973, p.167). Intuitively, however, it might be suspected that the kinds of questions teachers ask in an art class may differ from those raised in a physics class.

The need to devise instruments responsive to different subjects has been supported by Borich and

Klinzing who refer to "the significant influence of subject matter on the stability of the behaviors measured" (1984, p.40), and also by Stodolsky, who affirms that

> because subject matter is such a profound determiner of classroom instructional practice, it should be a primary contextual variable in any study of instruction or assessment of teaching (1984, p.15).

Language teaching researchers have long asserted their independence from research relating to content subjects (such as maths and physics) in view of the fact that in the language classroom, language is both the medium and the content.

Stern judges that "language teaching research has certain specific characteristics which make it different from other educational research because its subject matter is language" (1983, p.63). Corder (1968) points out that language is a universal feature of human behaviour, which cannot be said of history, for example; he recalls a distinction between 'performative' knowledge and 'cognitive' knowledge and takes the view that

> the job of language teaching is getting the learner to develop a performative knowledge of the language through the intermediary of both the performative and cognitive knowledge of the teacher (1968, p.79).

Long (1984) refers to a "sense that language classrooms differ in some fundamental ways from content classrooms" (p.419) and adds that "in most second language classrooms, language is both the vehicle and the object of instruction" (p.419). (See also Jarvis and Adams 1979,

p.3).

Given the widespread agreement about the special case of the language classroom, the implications for systematic observation are obvious: in Stern's words

> if a study requires classroom observation, the investigator can obviously draw on the experience in classroom observation that is available in educational research; but the categories that have been developed may have to be rethought to meet the conditions of the language class (1983, p.423).

That this rethinking has not always been done by FL researchers will become apparent.

While educational researchers have been using observation instruments for many years (Bales [1951] can be regarded as a forerunner of Flanders [1960]), language teaching researchers have been slower adopt the methodology. However, by now at least 29 instruments have been developed specifically for FL teaching. Most of these are covered in the 3 major reviews of the field: Mitchell (1977), Long (1980) and Allen et al (1983). The instruments are: Carton (1966); McArdle and Scebold (1968); Jarvis (1968); Nearhoof (1969); Wragg (1970); Rothfarb (1970); Moskowitz (1971 and 1976); Krumm (1973); Capelle et al (n.d.; a and b); Prokop (1974); McFarlane (1975); Delamont (1976); Freudenstein (1976); McEwan (1976); Long et al (1976); Allwright (1977); Seliger (1977); Fanselow (1977a); Politzer (1977); Riley (1977); Wesche (1977); Barkman (1978); Naiman et al (1978); Bialystok et al (1979); Mitchell et al (1981); Ullmann

and Geva (1982); Johnson et al (1985); and Frohlich et al (1985).

Most of these systems have been applied in only very small-scale ventures. Mitchell et al (1981) consider Moskowitz' (1976) study of the FLint system to be the largest - 22 teachers were observed for 4 periods each. In view of the fact that the most widely used educational systems, FIAC (Flanders 1960) and OSCAR (Medley and Mitzel 1958), have been used extensively for over 25 years by a large number of researchers, it is clear that FL instrument development is in its infancy.

Nevertheless, from tentative beginnings, in which there was wholesale borrowing, more elaborate work has been done in recent years (e.g. by Frohlich et al 1985).

The early borrowing indicates that the perceived differences between language and content classrooms failed to influence some of the pioneer FL systems. Moskowitz (1971) slightly elaborated on Flanders' (1960) 10 categories, adding a small number of items and redefining others, partly to enhance the recognition of affective characteristics (e.g. 'jokes' and 'laughter'). (For a critique of interaction analysis in general and of FLint in particular, see Bailey 1975). Wragg (1970) simply accepts the 10 FIAC categories without any modification whatsoever except to double them up so as to allow each category to be coded for 'native' and 'foreign' language use. As Mitchell et al note, one consequence of this is that "a coding of 10 stands in

this system for 'Silence in L1', a coding of 20 for 'Silence in FL'" (1981, p.5). In addition to Moskowitz (1971) and Wragg (1970), McArdle and Scebold (1968) and Rothfarb (1970) were also based on FIAC. As Long reflects, "it is surprising that so much borrowing should have taken place when one considers that second language classrooms differ from most others" (1983, p.9).

Other researchers developed systems that reflected FL concerns more consciously. Jarvis (1968), for example, distinguishes between 'real' language use and drill activity, and shows as much interest in student behaviour as in teacher behaviour (13 teacher categories and 9 student categories). Discourse analysis research has influenced some instruments (e.g. Allwright 1977). Fanselow (1977a), although his basic unit of analysis (the 'move') derives from Bellack et al (1966), is also explicitly concerned with FL behaviours. He distinguishes between meaning, grammatical and phonological content, for example, and between linguistic and non-linguistic media. What these instruments have in common is that they are at least to some extent based on a perception of the process of language learning.

More recently, researchers have been increasingly explicit in this respect. Mitchell et al, for instance, propose specific FL criteria:

> 1. Any system should be based on current theoretical understandings of the process of foreign language learning.
> 2. Systems should allow for multidimensional

> coding of lesson discourse, in terms of as many dimensions as appear significant on the basis of the theoretical understandings mentioned in 1.
> 3. Accepting the content of discourse as a phenomenon with its own internal structure, analysis systems should seek to preserve this structure as far as possible, adopting discourse units at one or more levels as the basic unit(s) of analysis, rather than time-based or other non-analytic units (1981, p.10).

Similarly, in their Communicative Orientation of Language Teaching (COLT) instrument, Frohlich et al state that

> Part A ... contains categories derived primarily from pedagogic issues in the communicative language teaching literature, and Part B ... reflects issues in first and second language acquisition research (1985, p.29).

One final consideration in this review of systematic observation is whether or not an instrument especially developed for FL classrooms is relevant to all FL classrooms. Allwright judges that "researchers may be such 'prima donnas' that they cannot bear to use anyone else's observation instruments" (1983, p.198). However, given that very few FL instruments are elaborate, it is quite possible that the needs of a particular investigation may demand at least modification of existing systems.

2.3.3 Standardisation of Treatment

An important question with regard to implementation is whether or not the treatment is uniform. This is relevant to an evaluation of the Bangalore project from two perspectives: firstly, it would be germane to know

whether the methodology proposed by Prabhu (see chapter 3) was standardised or whether it was allowed to vary naturally? secondly, if the CTP is adopted elsewhere, is it possible or desirable for it to be a replica or is it likely to be only loosely based on the original idea? and what factors influence this?

We have seen in chapter 1 that treatments in FL programs have often been inadequately monitored. The result is that treatments may have varied so widely as to be indistinguishable from the comparison treatments. One teacher's Natural Approach may be another's Cognitive Code. Therefore, it might be asked: is it desirable to make provision for standardised treatments or is it the variation that is of interest?

In this section, first of all, the notion of standardisation is elaborated (2.3.3.1); secondly, the extent to which it is feasible is questioned (2.3.3.2) as is the extent to which it is desirable.

2.3.3.1 Fidelity Approaches

There is a tradition which derives from the Fisherian view of experimentation (Fisher 1966) and endorsed recently by Cook and Campbell (1979) that the researcher should strive to ensure that treatments are uniform. This is so crucial for internal validity and for the objectivity of a study that some commentators have been moved to the use of evocative metaphors to make the point. Freeman (1964) contends that once the treatment

has been specified, the evaluator

> must continue to remain within the environment, like a snarling watchdog ready to fight alterations in program and procedures that could render his evaluation efforts useless (p.194).

He also talks of the need to "exercise sanctions to prevent slippage" (p.194).

In the educational literature, there have been many attempts to determine 'fidelity' - the congruence of specified treatment and implemented treatment. In Hall and Loucks (1977), the Levels of Use are based on the specified treatment; the closer the teacher adheres to the specifications, the higher the Level of Use. As Fullan (1983) points out, there is a difficulty here in that it may be easier to arrive at specifications of a program if it is highly structured; for Fullan, this is not a useful state of affairs because (as with program objectives) specification may not always be possible in a form that is assessable. Fullan and Pomfret (1977) offer an example:

> it appears as if some dimensions of implementation are more difficult to assess than others. For example, in the study by Gross et al [1971] 'try to act as a catalyst between children' may be much more subject to error in assessment than 'permit student interaction' (p.365).

Freeman's 'snarling watchdog' may not know a catalyst when it sees one.

However, teachers who are to implement the program may interpret acting as catalysts quite differently. In fact, Gross et al (1971) found that most of the teachers

in their study were unable even to identify the major features of the program they were using. Charters and Pellegrin (1973), Crowther (1972), Downey et al (1975), Lukas and Wohleb (1973) and Naumann-Etienne (1974) all found that low explicitness of treatments led to confusion and low implementation (cf. Fullan and Pomfret 1977, p.368-369).

The fidelity approaches to measuring implementation all assume that different adaptations by different teachers are aberrations. Churchman (1979) criticises the idea that "differences among teachers ultimately must be eliminated" (p.25). An extreme form of this attitude to standardisation is to be found in Alkin (1969). Alkin was attritional in his pursuit of stable behaviour, envisaging an implementation scheme which would require teachers to keep on repeating and modifying a treatment until uniformity had been achieved.

Cook and Campbell (1979), aware of the limitations of insisting on a single standardised treatment, recommend a factorial design involving planned variation. An example of this would be to try out certain varieties of the program in inauspicious circumstances and in an environment perceived as favourable, varieties with adult learners, varieties with children, and so on. But the varieties would still be closely specified and uniformity within them demanded. The advantage of this procedure is that more is known about more possible permutations of a program in more settings with more kinds of learner.

There are two basic objections to it: first of all, it may not be possible to talk realistically of replication; secondly, any form of standardisation, including those that make provision for controlled variation, do not address the issue of what happens when the evaluators have gone away and the pressure to conform is lifted (and thus whether or not standardisation is desirable). We will discuss both of these objections in turn.

2.3.3.2 The feasibility of Standardisation

With regard to achieving uniformity both within and between teachers, the idea that an educational program can be replicated owes more to rhetoric than practicability. When researchers in the behavioural sciences insist on a meticulous description of research procedures and of treatment operations, it is so that others may then replicate their study. In education, replication is rare, so the main function of meticulous description is to enable other researchers reading a report to judge the reproducibility of a study. In other words, reproducibility depends on a thought-experiment (Cronbach 1982, p.121-122).

If we insist on a program giving explicit specifications of itself, we are demanding that it should be reproducible, i.e. that different teachers will be able to understand what is required and carry it out uniformly. But the reproducibility of a treatment can be

considered at different levels of fidelity. If the specifications of an FL program, for example, state that all structural errors must be corrected, it may be that all teachers are seen to do so; however, each teacher may have corrected in entirely different ways. Even if a program managed to specify procedures so minutely that sources of variation were largely precluded, just one question from a student could entail a teacher's response that would differ from a response in a classroom where the same question went unasked.

Thus, uniformity of treatment is always a matter of degree. It can only be approximate. The question then becomes: how much variation is permitted before it is conceded that the treatment was not standardised?

For example, was the experimental program in the Harvard Project Physics (Welch and Walberg 1972) standardised? Welch (1983), pondering this a decade later, recalls discovering at the end of the project that one of the experimental classes had used as their major text the major text of the control group. He says: "I have often wondered about the contamination problems we missed" (1983, p.100).

Shaver (1983) considers that the current modus operandi in monitoring implementation is to make gross post hoc judgments about whether or not a program was standardised, and whether or not it could be distinguished from a control program. He cites his own (1964) study as an example: 15 critical differences

between two programs were specified; only 10 of these 15 were significantly different. Nevertheless, his conclusion at the time was to state that the teachers did in fact succeed in differentiating the 2 programs. Self-analytically, he has this to say:

> What had not been specified on an a priori basis was whether it was essential that all 15 hypotheses be confirmed or, if not, what confirmation would be acceptable, that is, which categories of behavior were especially critical to the differentiation of teaching styles, what magnitudes of behavioral differences were essential within categories, and what deviations from predictions for style differences could be tolerated across categories (Shaver 1983, p.7).

Since Shaver's article, there has been at least one attempt to specify levels of fidelity a priori. Following Hess and Buckholdt (1974), Wang et al (1984) set out to divide teachers into high, average and low implementors. They identified 12 critical dimensions of their program and, using percentages, set criterial levels:

> classrooms with scores at or above the 85% criterion level in 11 or 12 of the critical dimensions are identified as being at the high degree of implementation level; classrooms with scores at or above the 85% criterion level in 6 through 10 of the critical dimensions are at the average level; and classrooms with scores at or above the 85% criterion level in five or fewer critical dimensions are at the low-level (1984, p.268).

These prespecifications are, however, entirely arbitrary. Thus Wang et al's (1984) 'ad hocery' takes the place of Shaver's (1964) 'post hocery'. As Shaver (1983, p.8) was aware, there is simply no principled means available for specifying essential levels of implementation.

Therefore, when we talk of standardisation in educational programs, we are appealing to relatively unrefined concepts. Standardisation can only be 'enforced' in a rough and ready way, and assessed imprecisely and arbitrarily.

2.3.3.3 The desirability of Standardisation

A considerable level of interference may be necessary to have the effect of ensuring reasonable uniformity. If it could be assumed that for teachers to implement uniformly it was only necessary for the program plans to be mailed to each of them, then the resulting program would be high in both internal and external validity. That is, not only could we be sure that program X was actually implemented as planned, but also that it could be implemented in a variety of sites without Freeman's (1964) snarling watchdog to prevent alterations. Such a program would travel well; it would operate both within the confines of a controlled study and beyond. However, such a program may not be a realistic possibility.

Most of the literature on general factors that influence implementation stress the need for considerable intervention, especially on a personal level, so as to engage teachers' commitment and develop their sense of 'ownership' of the program (e.g. McLaughlin and Berman 1975; Crandall et al 1982; but see chapter 5 for a fuller account of this literature). These studies tend to

discourage optimism that intervention by program developers or evaluators need only be minimal. However, there are 3 studies which have specifically sought to test (among other things) the implementation that results from minimal intervention; these are Anderson, Evertson and Brophy (1979), Crawford et al (1978), and Good and Grouws (1979).

Interventions in the Anderson et al (1979) study consisted of researchers meeting the teachers, a manual being given to the teachers which was to be read before a follow-up meeting; then in a year-long period, most of the teachers were observed about 15 to 20 times. Teachers were found to implement so successfully that their students' achievement was significantly higher than for a control group. In the Crawford et al (1978) study, there were 3 groups: an observation only group, a minimal training (weekly manuals) plus observation group, and a maximal training (weekly manuals and weekly review meetings) plus observation group. All teachers were observed before, during and after the 5-week program. The minimally-trained group implemented as well as (in fact, slightly better than) the maximally-trained group. Good and Grouws (1979) had an introductory meeting with teachers, followed by a 90-minute period to describe the program behaviours, the distribution of the manual, another 90-minute meeting 2 weeks into the project; most teachers were observed about 6 times in the 4-month

duration of the study. Again, implementation was found to be satifactorily effected.

Good (1979, p.57), discussing these studies, tentatively concluded that

> elaborate delivery systems may not be necessary for effectively training in-service teachers to perform specifically identified classroom behaviors and (...) observation of teachers does not necessarily have to be a part of the in-service training.

This would seem to offer hope for minimal interventions, but, as Coladarci and Gage (1984, p.542) point out, none of the studies was minimal with respect to both training and observation. Coladarci and Gage set out to test the effect on implementation of a program that was minimal in both respects.

Manuals were mailed to teachers but the researchers never made personal contact. Observation was confined to a maximum of two 2-hour periods in the Fall semester and two 2-hour periods in the Spring semester. It was found that the treatment was hardly implemented atall. Coladarci and Gage conjecture that personal contact ruled out the possibility of conveying enthusiasm and gaining teacher commitment; further, they consider that observation in the 3 studies supporting minimal intervention may have had an 'enforcing' effect:

> while the manifest function of classroom observations is to obtain information concerning classroom characteristics and events, the latent function of such observations may be to facilitate treatment implementation (Coladarci and Gage 1984, p.550).

They note that teachers may unwittingly have come to

"regard the relatively frequent and lengthy classroom observations as a kind of supervision or monitoring" (1984, p.550). (This is particularly relevant to the Bangalore inquiry as there was a great deal of observation; see chapter 5).

Therefore, it would appear that the more we succeed in standardising a program, the more we reduce external validity because the greater the control needed to sustain appropriate behaviour. That is to say, while the pressure to conform is present, a degree of uniformity may be achieved, but once that pressure is lifted, or the program is implemented elsewhere, program realisations may bear little resemblance to the original conception.

So far, only attempts to standardise have been discussed, but the evaluator may be in a position to choose anything in a range from a study that attempts to be fully reproducible to one that allows program plans to be put into operation by local practitioners without exercising sanctions. Natural variation may either be considered "noise in the system" or "a phenomenon of interest" (Cronbach 1982, p.263). The main advantage of natural variation is that it enhances external validity insofar as information becomes available about what happens to a program under natural operating conditions. (To know this, of course, implementation would still need to be monitored).

Should a program fail to be operationalised

uniformly either within or between teachers, a researcher interested in natural variation would regard this knowledge as a finding and not as an aberration.

A study which insists on standardisation is artificial because in the real world, teachers adapt programs to meet local needs, to conform with their own perceptions of teaching, and so on. They do not adopt unless they are committed to a project or there is other pressure to comply. If what is of interest is what will happen to a program when it reaches the field, then standardisation of treatments would work against that goal.

In psychological research, some researchers have recommended "flexible administration of the independent variable [i.e. of the treatment]" (Carlsmith, Ellsworth and Aronson 1976, p.157), but they are aware that they are flying in the face of most teachings on research methodology: "we anticipate that many experimenters will disagree with us, suggesting that standardisation is the hallmark of an experiment" (p.157).

They are not alone. The Stanford Evaluation Consortium (a group of specialists from a number of diverse disciplines, including education, sociology, psychology, statistics and communication research) add their authoritative voice to the argument:

> allowing natural variation to occur and then appraising its extent makes interpretation comparatively easy. If findings are consistent from site to site, the PSC [policy shaping community] learns that the treatment has much

> the same consequences wherever and however it is installed. Insofar as the results differ, something much more important is learned: not all realizations that come under the same general label work the same way, and a plan to establish a uniform program by a centralized decision may be a fantasy (Cronbach and Associates 1980, p.277).

The merit of seeking information about natural variation, then, is that it points the way to what may happen in future programs and in a variety of circumstances.

2.3.4 Discussion

This section has reviewed the literature arguing the need to monitor implementation and also discussed the strategies (particularly systematic observation) which can satisfy this need. In addition, it has considered the role of standardisation of treatments - the feasibility and desirability of arranging for uniform realisation of program plans.

The view that emerges is that while implementation certainly needs to be monitored, normative implications for program realisation in the hands of different teachers and in different sites would have the effect of inhibiting external validity. This is because we would be setting up, for experimental purposes, what could not be maintained when the evaluator leaves the scene.

A program that achieves a high level of uniformity appeals to internal validity while a program that permits natural variation speaks to external validity. This raises the question of which kind of validity has a

principal claim on a program evaluator. The issue of the relative value of the two forms of validity now needs to be discussed, along with its corollary, whether studies are more usefully conducted in the 'field' or in the 'laboratory'.

2.4 Field-Studies versus Laboratory-Studies

First of all, it should be made clear that what is meant here by 'field-study' is long-term, classroom-based inquiry into the effect of relatively unstandardised complete programs; by 'laboratory-study', it is not supposed that the study is carried out in a laboratory - though some actually do take place in language laboratories - but that the study is short-term and only involves the testing of individual components in a theory in an environment in which extraneous variables are artificially held constant. These definitions allow some degree of latitude, but conform to the way the terms are widely used in the educational literature.

Whenever the studies by Scherer and Wertheimer (1964) and Smith (1970) are referred to in the applied linguistics literature, attention is usually drawn to what is widely perceived as a false dawn in language teaching research methodology, that is, the unfulfilled expectations of comparative field-studies (Freedman 1971; Allwright 1972; Stern 1983). Scherer and Wertheimer proposed a "rigidly controlled large-scale scientific experiment which would yield clear-cut data" (1964,

p.12). That this expectation should now seem so very sanguine is an indication of the extent to which the subsequent disappointments have become an established part of our lore.

As already noted (in chapter 1), the principal difficulty of field-studies has been the lack of documentation of the independent variable and the apparent impossibility of arriving at program-fair procedures for testing. Section 2.2, above, suggests that there are grounds for pessimism with regard to program-fair testing, which would raise scruples about the future of comparative studies; but field-studies do not need to be comparative (section 2.1.3 above), and the implementation literature (section 2.3 above) indicates that treatments can be adequately monitored, which argues that field-study is at least feasible.

The question, then, is not so much the viability of field-study but why we might wish to conduct this kind of inquiry at all. In putting a case for the desirability of field-study, the limited reach of laboratory-study is discussed in terms of internal and external validity (2.4.1.1), overall strategy (2.4.1.2), and its relationship with what happens in the field (2.4.1.3). It is argued that laboratory-study seeks understanding that will not necessarily have direct implications for continuing or future programs (2.4.1.4), whereas the priority in program evaluation is to determine what works

in real, that is, field settings (2.4.2 and 2.4.3).

2.4.1 The Limitations of Laboratory-Study

2.4.1.1 Internal versus External Validity

Central to the any question of evaluation design are the concepts of internal and external validity. To be reliable, the internal validity of a study must be high and to be usable so must its external validity. The difficulty is, as Campbell and Stanley (1963) point out, is that the two forms of validity appear to be in conflict, the demands of one militating against the demands of the other.

According to Campbell and Stanley (1963), internal validity has to do with factors that might constitute competing explanations for observed outcomes (i.e. something other than the treatment may have been responsible for obtained effects). Thus we might say that a study is internally valid (or that laboratory conditions pertain) if subjects are randomly sampled, treatments are standardised, and no extraneous influences intrude. (For a full account of the factors that jeopardise internal validity, see chapter 4, section 4.4.1).

Campbell and Stanley's (1963) view of external validity is that it is concerned with generalisability. Their treatment of this concept is slight relative to their attention to matters affecting internal validity. As we shall see (in chapter 4, section 4.4.2), the

concept is elaborated considerably by Bracht and Glass (1968), Snow (1974) and Cronbach (1982).

The conflict between the validities stems from the fact that the controls that are needed to ensure that a study is internally valid cannot say what will happen in circumstances where the controls are not imposed (i.e. in all real-world situations). While one might accept that all research involves a trade-off between the validities, it is far from clear that evaluation is best served by the primacy accorded to internal validity, especially in commentaries on language-teaching program evaluation: Long (1984), for instance, lists a number of threats to internal validity and neglects even to mention external validity.

Campbell and Stanley's (1963) conception of internal validity has held most researchers in thrall for many years. Hatch and Farhady (1982), like Long (1984), give it prominence and suppress concern with external validity. Very few have questioned just how limited the reach of the conception of internal validity (of Campbell and his various co-authors) is, but Cronbach (1982) has put forward coherent arguments which may help to draw attention to this.

Campbell and Stanley say that the question which addresses internal validity is: "Did in fact the experimental treatments make a difference in this specific experimental instance?" (1963, p.175). Cronbach argues that this is probably interpreted by most readers

as meaning that a class of treatments, e.g. a method with a label such as 'suggestopedia', caused a difference. In fact, however, the use of the past tense and of the phrase 'in this specific experimental instance' indicate that this is not what he meant. On the contrary,

> they define internal validity as pertinent only to an interpretation of a particular historical event. The interpretation is not a prediction about other instances, not a lawlike statement. (Cronbach 1982, p.127).

Cronbach contends that Campbell has always meant the term 'internal validity' to refer to "an inference devoid of generalisation" (Cronbach 1982, p.128). That is, the only conclusion capable of having internal validity is that something made a difference. Labeling the cause is not part of Campbell's claim for internal validity (Cook and Campbell 1976; 1979). What this means is that it is not even possible to say that a specific realisation of suggestopedia made a difference, because it may simply have been a question of teacher motivation. Thus, as Cronbach states, "Campbell's 'made a difference' inference is minimal" (1982, p.128).

If any factor associated with a past treatment could have caused a difference, then the term 'causality' is hardly a useful one:

> I consider it pointless to speak of causes when all that can be validly meant by reference to a cause in a particular instance is that, on one trial of a partially specified manipulation t [treatment] under conditions A, B, and C, along with other conditions not named, phenomenon P was observed. To introduce the word cause seems pointless. Campbell's writings make internal

> validity a property of trivial, past-tense, and local statements (Cronbach 1982, p.137).

If, as Cronbach argues, internal validity is so limited, the question for the evaluator is: in a trade-off between the validities, would it be sensible to sacrifice arrangements that would enhance external inference for controls that would support a trivial, past tense and local interpretation of a historical event? Surely, all evaluators would wish to generalise their findings, so it would seem that conceptions of external validity must now be examined.

It was mentioned earlier that the Campbell and Stanley (1963) conception of external validity was a matter of generalisability:

> External validity asks the question of generaliability: To what populations, settings, treatment variables, and measurement variables can this effect be generalised? (p.175).

They listed 4 types of threat to external validity. These were later extended by Bracht and Glass (1968) and Snow (1974) (see chapter 4, section 4.4.2), but it was Cronbach who gave precedence to external inference, arguing that "internal validity is of only secondary concern to the evaluator" (1982, p.112).

Where Cook and Campbell (1979) limit internal generalisation to persons - from the sample of subjects to the population from which they were drawn, Cronbach (1982, p.116) would insist that sampling of treatments and tests should also be planned. For Cronbach, this statistical conclusion validity is within the realm of

internal inference. For him, the wish to make a statement about a population, treatment or set of testing arrangements that was not systematically represented in the study involves a generalisation which he prefers to call an "extrapolation" (1982, p.119). Unlike Cook and Campbell (1976, 1979), Cronbach includes construct validity within external inference, that is, he takes the view that external inference involves making predictions on the basis of known differences between the situation that was studied and another situation (1982, p.119-120).

To support such extrapolation, Cronbach recommends reducing the differences that must be allowed for and backing the inference with supplementary information. Thus, if in the planning stage of an evaluation, the questions that others will want answered are specifically addressed, the distance that the extrapolation must travel is shortened and the inference is rendered more credible. Supplementary information could be derived opportunistically from any suitable source: "folklore, history, anecdotes, research on tangential topics" (Cronbach 1982, p.290).

Perhaps the conflict between internal and external validity and the differences between Campbell and Cronbach can best be summed up as follows: Campbell and Stanley (1963, p.175) state that "internal validity is the sine qua non", Cronbach (1982, p.114) avers that "relevance is surely the sine qua non in evaluation".

In the language-teaching studies reviewed (in chapter 1), many researchers opted for internal validity at the expense of external validity. Discussing the tension between the two validities, Freedman flatly states that "one has no alternative but to choose the lesser of the two evils" (1982, p.132). For her, the lesser evil is internal validity. Reviewing the failures of a (very) few global, field studies to achieve the tight control that would promote internal validity, she concludes that "the road from Pennsylvania [i.e. Smith's 1970 study] must surely, therefore, now go in the direction of a series of small-scale experiments" (1971, p.37). (Freedman's use of the terms 'large-scale' and 'small-scale' roughly correspond to my use of the terms 'field' and 'laboratory' study; Freedman 1979, p.187-188). She thinks there is no point in even considering field research any further: "it is not a question of finding ways to control the variables in large-scale experiments, since it is the very 'size' or global-ness of the experiment which precludes rigid control" (1971, p.36). The Pennsylvania project, she submits, can be seen as a field-study that was "controlled as far as it was possible to do so" (1971, p.36). She is quite explicit: field-study cannot be controlled sufficiently; therefore external validity must be sacrificed for internal validity; consequently, what is needed is laboratory inquiry.

Less boldly proclaimed, perhaps, this is, neverthe-

less, essentially the view taken by many language teaching researchers. Thouless suggests that instead of overall method comparisons, a more profitable line of inquiry would be to "make a study of a single element in a new approach to foreign language learning" (1969, p.219) in order to measure its contribution to learning efficiency. In much the same vein, Hammerly comments that large-scope experiments shed little light and that the "only possibility for fruitful research in our field lies in ... small-scope experiments" (1982, p.638). Levin points in the same direction, arguing that "if specific variables are selected for study ... there is a good possibility that research will prove parts of each theory to contribute to methodological advancement" (1972, p.40). Davies advocates research operations that will establish a "satisfactory set of procedures within an overall theoretical approach" (1977, p.1). Carroll, too, agrees, observing that "a theory implies an interconnected set of hypotheses, each of which can be tested in a separate experiment" (1965, p.280). Citing Carroll with approval, Seliger (1975, p.10) expresses a preference for laboratory study, as do Hocking (1969), Von Elek and Oskarsson (1973) and many others.

It would appear then that there has been something of a consensus that field-study has proved barren, that laboratory studies should be pursued, and that internal validity is paramount.

2.4.1.2 The Overall Strategy of Laboratory Study

While there is considerable agreement on the need for laboratory inquiry, there is less obvious concurrence of ideas concerning the overall strategy of such studies, or the framework within which they might operate. Carroll (1965), Levin (1972) and Davies (1977) seem to agree that a theory is divisible and that each component can be tested individually. Similarly, Freedman (1976, p.25) envisages a series of specific comparisons of method components contributing to a composite whole.

Lado (reported in Smith 1970, p.349) seems to take a rather different view of strategy:

> if in a test-tube experiment one finds out that the lab used in a certain fashion <u>does</u> produce something - then later you find out in a massive experiment that it is not producing - then one knows <u>why</u> it is not producing. If the large scale is done first, one cannot isolate contributing factors.

He sees laboratory experiment as a necessary preliminary to a field venture. Brumfit, too, though he does emphasise the importance of field inquiry, contemplates the same progression from laboratory to field investigation:

> The researcher should be concerned with providing evidence for real changes in typical situations, by applying the results of small, controlled experiments to large-scale but typically based experiments (1980, p.136).

It would seem that there are two fundamental views of strategy. The first, subscribed to by Carroll (1965), Levin (1972), Freedman (1976) and Davies (1977), is that

a series of laboratory studies will amount to a comprehensive assessment of a theory, as in Figure 2.1, below. The second, endorsed by Lado (reported in Smith 1970) and Brumfit (1980), is that a field study will follow a laboratory study, and in this way, presumably, indicate the validity of a theory. This can be summarised as in Figure 2.2, below.

Figure 2.1

Strategy for laboratory studies: testing each component of a theory

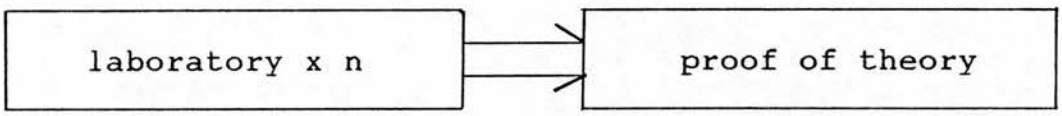
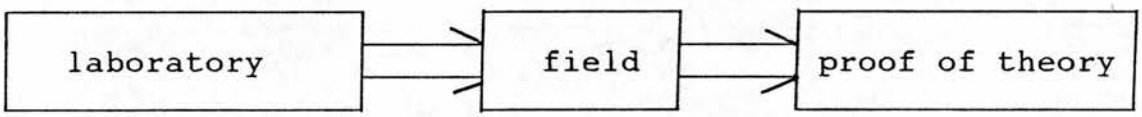


Figure 2.2

Strategy for laboratory studies: from laboratory to field setting



2.1.4.3 The Relationship between Laboratory and Field

Both of the perspectives condensed in Figures 2.1 and 2.2 are based on assumptions which, it might be argued, are untenable. The first view is premised on the belief that the sum of the parts will equal the whole. The second view assumes that what happens in an artificial setting has a knowable relationship with what happens in a natural setting. Let us suppose that a

laboratory inquiry fails to show a significant difference between the issues under study. According to the first view, the theory is deficient in a constituent part. According to the second view, there is probably no point in proceeding to a field-study forum. Alternatively, if a laboratory inquiry demonstrates that one factor has a significant effect on learning, the first view would claim that part of the overall theory has been borne out. The second view would sanction the instigation of a field investigation.

However, what both views ignore is that each laboratory study is, in Weiss' words, "the prisoner of its setting" (1972, p.79). This is true whether a series of interlocking experiments is being considered (as in the first view) or whether manipulated experiment is to serve as a basis for embarking on a field exploration (as in the second view). Both views ignore the 'synergy' factor, that is, the myriad interaction effects of separate elements that distinguish a laboratory environment from a typical setting. In other words, both views ignore external validity.

That the assumptions these views are based on are false has been well documented in the educational literature. Good and Power report that "experimental research based on one or two stimulus events ... frequently yields positive results that do not hold up in a naturalistic setting" (1976, p.47), and cite, as an

example, Kounin's (1970) finding that teachers' desist techniques directed towards individual students had a ripple effect on classmates in a laboratory setting, but no effect whatsoever in a naturalistic setting.

Another example indicating that theoretical principles derived from laboratory studies do not hold up in a field context is the Plaza Sesamo evaluation in Mexico (Diaz-Guerrero et al 1976). Phillips (1981), commenting on this study, points out that in the laboratory setting, the series of television programs produced clear gains in children who watched them, but that in the field-trials, the gains never materialised. He offers the opinion that

> real problems emerge when the results of the experiment are applied outside the controlled research setting ... in a laboratory experiment a factor might produce an effect, but in the field it might fail to do so (Phillips 1981, p.18).

Barker (1965) distinguished between psychologists as transducers (non-manipulators in a field-setting) and operators (manipulators in a laboratory setting). His investigation into frustration in children (Barker et al 1941) was extended by his student, Fawl (1963), "from children in vitro, so to speak, to children in situ" (Barker 1965, p.5). Fawl found that when frustration occurred, it did not have the behavioural consequences observed in the laboratory. Barker considered that the data which psychologists produce as operators and transducers "refer to non-overlapping classes of

psychological phenomena" (1965, p.4), and wondered

> how the properties which behavior units possess when they are lined up Indian file by an operator are modified when they occur in overlapping formation, as they so often do in the phenomena reported by T, i.e. transducer data (1965, p.7).

Cronbach speculates that a "laboratory generalisation, once achieved, may not be a good first approximation to real world relationships" (1975, p.21). Aware of this predicament, Brunswik (1955) and Snow (1974) recommend greater ecological representativeness in research. Hilgard and Bower (1966) made the point that the argument that more complex behaviours could be better understood once simple behaviours in simplified conditions were more fully understood is without research validation.

Finally, and perhaps more disturbing, is that "a factor which fails to produce an effect in the laboratory might work well in the field" (Phillips 1981, p.18). So far, the author has been unable to find any examples of this possibility, which may be negative evidence for the wide currency of the assumption that if no effect is found in the laboratory, the issue is not thought to be worth pursuing in the field.

Laboratory study, since it cannot serve as a proxy for what happens in the field, may simply be investigating a different problem, which Raiffa (1968, p.264) and Mitroff and Featheringham (1974, p.383) refer to as "error of the third kind". This occurs whenever a

'wrong' inquiry is undertaken than would provide specious support for a hypothesis. The terminology has surfaced again more recently in the psychological literature: Doerfler and Chaplin readopt the notion of Type III error to illustrate the disparity between "laboratory analogues" and "the natural environment" (1985, p.227).

Many other researchers have expressed the same concern that the artificiality of laboratory inquiries may be unrelated to what would occur in typical settings (Orne 1962; Bannister 1966; Sells 1966; Cattell 1966; Allport 1968; Pereboom 1971; Harre and Secord 1972; Mitroff and Blankenship 1973; Babbie 1975; Gadlin and Ingle 1975; Mahoney 1978; Wachtel 1980).

2.4.1.4 The Function of Laboratory Inquiry

At this point, it may be worth stressing that although laboratory research may be unrelated to what happens in the field, this does not mean that there is no role for such inquiry.

Berkowitz and Donnerstein (1982) contend that laboratory studies need not be concerned with natural conditions. On the contrary, they celebrate artificiality, arguing that insofar as it derives from control over extraneous variables, "artificiality is the strength and not the weakness of experiments" (1982, p.256). In their view, the function of laboratory experiments is to test causal hypotheses, but not to "determine the probability that a certain event will occur in a particular

population" (1982, p.247). Mook is sympathetic to this perception, affirming that while laboratory inquiries cannot be used to predict real-life behaviour, they address the question of whether "something can happen rather than whether it typically does happen" (1983, p.382). He adds that "even where findings cannot possibly generalise and are not supposed to, they can contribute to an understanding of the process going on" (1983, p.382). To illustrate his point, he cites an experiment carried out by Argyle:

> targeted persons were judged at 13 points of I.Q. more intelligent when wearing spectacles and when seen for 15 seconds: however, if they were seen during five minutes of conversation, spectacles made no difference (Argyle 1969, p.135; cited in Mook 1983, p.382).

Argyle (1969) presents this as an indication of how isolating the independent variable can produce irrelevant or misleading results. Mook allows that from an applied perspective, Argyle is perfectly correct, but insists that knowing that the 15-second bias can occur is valuable, and asserts that the whole point of isolating variables that "come packaged in Nature is not necessarily to generalise back to the real world, but to increase understanding" (Mook 1983, p.384).

It is evident that there is a conception of laboratory research as a different kind of inquiry from field inquiry both in terms of methodological procedure and purpose. Essentially, they look for distinct kinds of information, the former affording insight into what is

possible, and the latter probing what is probable in certain populations. Laboratory inquiry is not undertaken in order to provide immediate feedback to practitioners, whereas field inquiry may be sufficiently relevant to motivate modification of teaching behaviour or to improve programs. (Although, exactly how research findings might be disseminated to teachers is a much debated issue and involves ethical considerations. For the process-product researchers, e.g. Gage [1978], findings will suggest a series of teacher behaviours that should then be adopted, while for others such prescription is anathema, e.g. Fenstermacher [1979, 1982]; Garrison and Macmillan [1984]. Fenstermacher [1979] advocates bridging research and practice by presenting research findings in such a way that teachers may conceptualise the issues in question without normative implications being conveyed).

Evaluation of educational programs, however, may not be an appropriate forum for consideration of what <u>can</u> occur. As Cronbach and Associates (1980) put it,

> For a technique of health care, a trial under highly controlled conditions almost always precedes a more realistic field test. An educational or welfare service is much less likely to be installed as a superrealisation [i.e. a manipulated study in ideal conditions]. Since the theory underlying an educational proposal is not so definite as the biological hypothesis from which a vaccine is derived, there is less interest in what is ideally possible (p. 239-240).

2.4.2 Evaluation as Field Study

To a large extent, the desirability of field-study in evaluation is justified by the argument that variables do not have the same effect in isolation as in combination, and that what happens in the laboratory may have little relation to what happens in the field, particularly as the two modes of inquiry do not seek answers to the same questions. It is justified because program evaluation intended to be of direct relevance to teachers, administrators, parents, in fact, all potential stakeholders in the community.

Talmage states that "program evaluation is an applied discipline, and it must confront the real world" (1982, p.595). Therefore, as Kratchowill observes, examining isolated variables in laboratory conditions "may not serve applied purposes" (1979, p.60). Also, Azrin's specifications are relevant:

> applied research has different requirements from basic research. Applied research emphasises outcome versus conceptual analysis ... situational complexity versus stimulus and laboratory simplicity (1977, p.141).

The orientation towards applied inquiry begs the question of what use is to be made of evaluations. In the past, it would appear that an all too familiar kind of use has been no use. For example, a series of studies extending over years and massively funded were reported in the 1970s; these studies compared the effectiveness of education in a number of subjects in a number of countries; one of them investigated the teaching of science in 19 countries. Here are some excerpts from the

conclusions:

> It must be confessed at the outset that limitations of time and money imposed serious constraints throughout the study ... The report may therefore perhaps be best regarded as a preliminary statement on the results of a vast undertaking (Comber and Keeves 1973, p.286).

Comber and Keeves talk of "differences which merit further investigation" (1973, p.299) and punctuate their closing paragraphs with caveats like "difficulties in the data preclude a definite answer to this question" (p.300). In short, "the present findings, their virtues and their faults, will serve above all as a point of departure" (p.300).

Needless to say, the 'point of departure' proved to be the end of the journey and the 'preliminary statement' pronounced the last words on the subject. There can, of course, be no complaint about the candour of the researchers' tentative conclusions; it might be asked, however, whether the expenditure of time, money and effort was in proportion to the usefulness of the findings. The evaluation did not set out to address answerable questions, and did not arrange data collection and analysis to provide usable information to researchers, administrators or teachers. The study was not oriented to providing for extrapolation.

Recently, the evaluation literature has focused on the utilisation of evaluations. A major voice in this new focus has been Patton's (1978). He argues that how information is to be used should dictate how the

evaluation is to be designed and conducted. King (1982) comments that nobody any longer believes that merely providing decision-makers with results of an evaluation will lead to use of those results. By contrast, evaluation must be user-focused (Henderson et al 1983). O'Keefe (1984) observes that

> several lessons can be learned from the evolution of federal evaluations. The first and foremost lesson is that evaluation will be useful only to the extent to which they relate to questions policymakers want answered (p.71-72).

He lists 4 other lessons. They are (a) "that the questions asked should have some chance of being answered" (p.72), (b) that evaluators and policymakers stick to their own areas of expertise, (c) that "evaluation is not the same as research" (p.72); by this O'Keefe means that tentative results hedged about with error probabilities do not assist policymakers; inconclusive findings can help the researcher develop stronger hypotheses but "they are useless to the policy-maker who wants help in making decisions that must be made today" (p.72). Leviton and Hughes (1981) also find that the most important variables affecting use are (a) relevance to the user and (b) credibility of the evaluation information to the user. Daillak (1983) argues too that evaluators must plan for evaluations to be usable.

As Cronbach and Associates (1980, p.3) state, "the distinction between evaluation and policy research is

disappearing". Even more strongly stated, Cronbach (1982, p.ix) asserts that "the logic of science must come to terms with the logic of politics". It appears to some critics of this political attitude towards evaluation that researchers are being asked to put aside their research tools and throw caution to the winds. For example, Bryk (1981), in reviewing a well-known educational evaluation report, notes that the report "is consistent with emerging canons for improving the utility of evaluation and policy research" (p.507) and chides the authors for failing to provide a disciplined inquiry. Bryk concludes that

> social scientists are faced with a dilemma. In striving to make applied social science research more useful for social policy, we run the risk of stripping the research process of its character ... While we may have figured out how to catch the policymaker's ear, we may have nothing special to say (1981, p.507).

The deliberate polarisation of 'research' and 'evaluation' by, among others, O'Keefe (1984) and Lincoln and Guba (1985) have perhaps given rise to unnecessary doubts. Of course, definitions of both can be arranged to reveal little overlap, but such definitions may be tactical, insofar as they seek to claim potent words for one side and to deny them to the other.

Ideals of parsimony and elegance would be less stressed in the kind of evaluation indicated by those concerned with utilisation, it is true, but any definition of evaluation insists on disciplined inquiry.

A description of disciplined inquiry is offered by Cronbach and Suppes (1969)(this is cited again in chapter 5, section 5.2.1):

> [the report of a disciplined inquiry] has a texture that displays the raw materials entering into the argument and the logical processes by which they were compressed and rearranged to make the conclusion credible (p.15-16).

This is a crucial property of any form of scholarly inquiry. It leaves the door open for both naturalistic and experimental inquiry and the tools approriate to each without permitting the charge levelled by Bryk that "the evidence is assembled to support a particular policy recommendation" (1981, p.507).

However, it would certainly appear to be the case that a concern for true experiment is suppressed by the need to provide usable information. Whether or not to set up a true experiment depends on how focused the research question is. What it boils down to is: do we know that the manipulations are precisely the ones of interest? For example, is a language teaching method ever sufficienly definite? The review of chapter 1 suggests that such notions are characterised by considerable vagueness in terms of actual practice. In view of this, it seems doubtful that there would be interest in what is <u>possible</u>, but rather in what tends to happen to methods when they reach the field. Methods, that is to say, are inexplicit and thus seem unlikely to contribute greatly to nomothetic knowledge; however, it may still be possible to illustrate how a program was implemented,

what processes were observed, what effects, and to attend to interactions that might explain variation. This would at least point the way to better future programs and guide potential implementors. Perhaps this is a more prudent aspiration for method inquiries.

Evaluators may ensure that they address relevant questions by the expedient of asking, like Rockwell (1982), three basic questions:

> 1. Who is interested in the evaluation and why?
> 2. What decisions are to be made as a result of the evaluation?
> 3. What do you want to know about the project?
>
> (Rockwell 1982, p.198).

If atall possible, it would be desirable to avoid the predicament that Abt and Magidson (1980) found themselves in: they collected a massive data base over 5 years, "yet, in retrospect, we still cannot adequately answer the most basic questions - 'what is the treatment?', 'who received it?' and 'why?'" (1980, p.208).

2.4.3 Discussion

This review, in attempting to clarify the evaluator's brief, has stopped just short of suggesting that language teaching program evaluation need no longer concern itself with theory. However, it is less diffidently proposed that what happens in practice should take priority.

Following the disappointments of field-studies such as Scherer and Wertheimer (1964), many researchers

turned to laboratory research as an option more likely to yield dependable knowledge with regard to foreign language pedagogy e.g. Levin (1972), Freedman (1976), Seliger (1975) and Wagner and Tilney (1983). All of these studies tell us what can happen in artificially conditioned environments.

On the other hand, these studies are externally invalid, and their findings cannot be generalised back to the classroom. They cannot say what is or what is not likely to work in particular schools, and therefore cannot be used as a basis for promoting or rejecting their adoption. They are simply not designed to; it is not their purpose. Field study, by contrast, can arrange to be of relevance to the ongoing programs and decisions to adopt programs.

It has been argued that user-oriented field study is desirable. The advantage of according it priority is clear; it fosters a primary concern for relevance rather than elegance.

Chapter 2 has provided a brief overview of educational evaluation; it has argued that the difficulties of program-fair testing raise scruples about comparing programs; also, it has demonstrated that programs need to be monitored and has described some of the principal means of doing so; finally, it has put a case for the precedence of external validity, encouraging heterogeneity of treatments, arguing for field study, and drawing attention to the requirements of potential users

of evaluation reports. In short, it makes the point strongly that field study is essential and that laboratory inquiry is probably misguided in evaluation. It is stressed above all that the main goal of an evaluation is to be relevant. The influence of this review and these views will be apparent in the 4 data-based chapters (4, 5, 6, and 7).

CHAPTER 3

A CRITICAL ANALYSIS OF THE BANGALORE PROJECT

3 A Critical Analysis of the Bangalore Project

This chapter describes and critically analyses the Bangalore project. The sources for the critique are basically of six kinds: (i) the considerable published literature that has accumulated around the project, (ii) the British Council reports of visitors to the project, (iii) the studies carried out by M.A. / M.Sc. students at the universities of Lancaster and Edinburgh, (iv) the RIE (Regional Institute of English, Bangalore) Newsletters and Bulletins, (v) a number of papers given at International conferences, and (vi) the author's visit to the project and numerous conversations with CTP teachers and the director, Dr. Prabhu.

The chapter begins with the description and follows with the critical analysis.

3.1 Description of the Bangalore Project

3.1.1 Brief Introduction

The Bangalore / Madras Communicational Teaching Project (CTP) has aroused a great deal of professional interest in recent years. It constitutes, in Howatt's (1984) term, a 'strong' form of a communicative curriculum. Initiated and developed by Dr. Prabhu in South India, in far from ideal circumstances and in actual school settings, it subjected a major current model of language learning - one that stresses the value

of unconscious assimilation - to a substantial trial which lasted for 5 years (1979 - 1984). In the following sections, the background to the CTP and its gradual consolidation are explored.

3.1.2 Background to the Project

The CTP began life as a local response to a local problem. It grew out of a sense of professional disenchantment with the prevalent structure-based approach that had failed to produce results satisfactory to Prabhu and a number of colleagues.

In the early 1950s, the first structural syllabus was introduced in South India. It was perceived as a major breakthrough as Forrester's (1954) account of the 'Madras' syllabus makes plain. However, several years later, it was observed that many teachers had merely applied grammar-translation methods to this new syllabus (Patel 1962; Smith 1962). That is to say, teachers tended to revert to procedures they were familiar with, so that the new syllabus constituted a case of what Stenhouse (1975) has called 'innovation without change'.

The subsequent campaign to rectify the situation has become almost legendary. The 'Madras Snowball', as it became known, was an attempt to retrain over 30,000 teachers of English according to structural principles (Smith 1962; 1968). First of all one group was thoroughly trained; each individual would then go into the field and train another group, and so on, in ever-expanding waves.

It was a massive undertaking and it is important to be aware of the scale of the investment in structural teaching in order to understand the apparent resentment within South India when the CTP began to attract interest in many quarters.

Although structural syllabuses and teaching methodologies were apparently implemented with considerable determination, by the late 1970s, dissatisfaction with them led Prabhu and his associates to cast about for alternatives that might promote more effective language learning.

Notional/functional syllabuses were considered at a seminar in Bangalore in 1978, but it was concluded that a change of syllabus content would be unhelpful. Such a change would merely substitute one set of objectives for another: control of the semantic structure of the language for control of its grammatical system. The project team believed from the beginning that the generative nature of grammatical structure was a powerful argument in favour of its centrality as a goal in language teaching (RIE Newsletter 1980, Vol.1. No.4). Although the goal seems never to have been in doubt, the question of how best to arrive at it remained to be elaborated. Thus an exploration of methodology commended itself to Prabhu and others more convincingly than did the adoption of an alternative syllabus.

The catalyst that gave direction to this

methodological exploration came at a seminar in Bangalore in 1979, chaired by H.G. Widdowson. Discussion of the use / usage distinction (Widdowson 1978) raised questions about the acquisition of usage itself, from which it emerged that usage might be enlarged by use-based language behaviour. In other words, the structure of the language might be acquired through communicative activity (see Prabhu 1981). This implied a basic methodological principle:

> not 'English for communication' but 'English through communication'; not 'learn English so that you will be able to do and say things later' but 'do and say things now so that as a result you will learn English' (Prabhu 1980a, p.23).

For this reason, the term 'communicational' rather than 'communicative' was preferred to characterise the project.

The guiding principle of the CTP was that form is best learned when the learner's attention is focused on meaning. To experiment with this principle, a series of activities was tried out in the classroom (RIE Newsletter 1979, Vol.1. No.1., p.10). A process of trial and error revealed that role-playing and dramatisation did not work; nor did the narration of stories without endings that the pupils were to complete. It was felt that the most promising results were obtained with tasks that involved problem-solving (it should be stressed that this was purely judgmental and not empirical). Consequently, tasks of this type came to dominate.

Given that the central tenet of the evolving methodology was that form is best learnt when the learner's attention is focused on meaning, it was argued that any syllabus based on a focus on form would directly conflict with it. It was argued that a syllabus that specified linguistic behaviour would make nonsense of a methodology that stipulated meaning-based activities (Prabhu 1980a; 1984). Thus the notion of a linguistic syllabus was abandoned in favour of a syllabus dictated by the methodology, and the CTP moved towards a task-oriented or 'procedural' syllabus (that is, it did not plan a syllabus but arrived at one that grew out of the methodology and classroom trial and error).

3.1.3 Development

The assumption that focusing on meaning facilitates the assimilation of structure implies that to a great extent language structure can be acquired and operated unconsciously. This perspective is hardly new (see for example Palmer 1921/1964). Perhaps the best known readoption of this view in recent years is found in Krashen's work (1981; 1982). However, unlike Krashen, Prabhu has not annexed the word 'learning' and narrowed its scope; he has not argued that what is learnt consciously cannot seep into the unconscious. Prabhu's reason for avoiding explicit attention to language is that it may conflict with the learner's constant process of hypothesis construction and revision, i.e. that there

is no reason to assume that the linguist's generalisations about language structure parallel whatever generalisations are actually involved in the learner's process of grammar construction (Prabhu 1982). Conscious learning based on linguists' generalisations is thought likely to be at odds with the learner's formative, transitional competence. In this sense, conscious learning is regarded as a hindrance to natural learning. By contrast, what is learned naturally is considered more readily available for deployment:

> In the Bangalore project ... we concentrate on the ability to communicate, and assume that the acquisition of structure will arise out of this, and, furthermore, that the kind of structure acquisition that arises from this is a better form of competence, is more dependable than the kind of structural competence that results from the teaching of structure itself (Prabhu 1980b, p.161).

Linguistic specification is therefore abjured. By the same token, no provision is made for pre-selection of language for any particular lesson, nor for classroom activities that focus on language. In their place, the project proposes reasonably challenging problem solving activities that are judged to promote understanding, and it is argued that this involves an incidental struggle with language use (Prabhu 1980a, 1981). The learner's attempts to cope as well as possible with the language required are thought to be essential to the process of grammar construction; this implies that the linguistic resources required for the completion of a task are best

perceived and internalised when the mind is engaged in analysing (Prabhu 1980a, 1981, 1982).

CTP lessons comprise 3 stages: (1) pre-task, (2) task and (3) feedback. The pre-task makes known the nature of the task, brings relevant language into play, regulates the difficulty level of the task, and allows some learners to learn from attempts made by others. The task itself is a period of self-reliant effort by each learner to achieve a clearly perceived goal (e.g. interpreting a schedule or a map). The feedback gives the learners an indication of their success on the task.

The pre-task is essentially a rehearsal during which the teacher's help is overt. The further the actual task is from the pre-task, the greater the challenge, and vice-versa. Achieving the appropriate regulation of the level of challenge is judged to be crucial to the success of the methodology. If the task is too difficult or too easy, interest flags and the learners' minds fail to become engaged. The criterion used by the CTP teachers is that at least half the class should succeed in performing at least half the task. One teacher stressed that if the level of the task was ill-suited to the class, "pandemonium" ensued (Bose 1980, p.86-87). (For a fuller description of the 3 stages of the lesson, see Prabhu 1980c).

It was mentioned earlier that the CTP arrived at a syllabus of tasks. A task-based syllabus was thought essential to ensure a sense of continuity and sequential

progression. In more concrete terms, it provided a loose framework in which each task could be translated into a lesson plan. The sequencing is partial: tasks dealing with the same general topic are divided into several cycles that recur intermittently throughout the course. A task-cycle consists of 3 to 5 sequential lessons on one theme. Teachers determined that if the cycle were any longer, learners suffered from what has been termed 'task-fatigue' (see, for example, Elia 1981, p.32).

Although, as has been discussed earlier, language content is said not to be predetermined, Prabhu notes that there is a certain kind of language control, that is, the language that forms the classroom input is controlled by the teacher. It is argued, however, that this does not necessarily imply that the content is predetermined. Much is made of the teacher's capacity to exercise 'natural' language control, that is to say, a continually modified, intuitive judgment as to what the learner can manage at any given stage. (This is distinguished from 'planned' language control, which <u>is</u> predetermined; see Prabhu 1982; 1987). The learner's mind engagement then makes the 'naturally' controlled input available for intake. It is not assumed that intake will be uniform in any way, so errors are not considered as deviant behaviour, but are dealt with incidentally and treated as contributory to the task at hand. Thus, incorrect language might be rephrased or ignored, but

there would be no attempt at generalisation, exemplification, or grammatical explanation (Prabhu 1982; 1987; see chapter 7).

There are three principal constraints in the CTP methodology: (i) low technology, (ii) reliance on reasoning, and (iii) reliance on teacher-class interaction.

'Low technology' simply means that since there are no more teaching aids available in typical South Indian classrooms than chalk-and-board and paper-and-pencil, that is all that the CTP demands.

'Reliance on reasoning' refers to the fact that CTP tasks involve reasoning, but not "personal feeling, preference or opinion" (Prabhu 1982, p.6). The reasons for this are that Prabhu maintains that learners "feel a sense of security in working with problems that have clearly right and wrong answers" (1982, p.6). Also it encourages guessing (considered desirable). A third reason given is that more open-ended problems make excessive demands on the learner's developing language (1982, p.6). Finally, it is contended that

> English is, in general, the language of rationality, rather than of emotion, for Indians (Prabhu 1982, p.6).

In regard to the 'reliance on teacher-class interaction', Prabhu ventures the opinion that CTP teaching "does not involve learner-learner interaction" (1982, p.6) firstly, because learners will probably revert to L1; secondly, because it would consitute a

break with tradition; and thirdly, because it might promote pidginisation (1982, pp.6-7).

So far, no comments have been offered on any aspect of the CTP. The opinions, arguments and convictions are taken from Prabhu's writings and addresses to international conferences. These views have provoked a variety of reactions which will now be set forth in the following critique (section 3.2).

3.2 Critique of the CTP

This critique includes an account of the professional excitement engendered by the CTP (section 3.2.1), four broad areas of controversy (sections 3.2.2 to section 3.2.5), and a report of Master's degree dissertations devoted to aspects of the project (section 3.2.6).

3.2.1 The Arousal of Professional Interest

The CTP quickly attracted attention. Roberts judged the project likely to "arouse considerable interest" (1982, p.190). Brumfit notes that "so many methodologists and applied linguists find Prabhu's approach to a procedural syllabus so exciting" (1984, p.240). Howatt is prepared to state that "whatever happens, Bangalore has set the context for one of the most interesting arguments of the eighties, if not beyond" (1984, p.288).

Prabhu, as a British Council English Language

Officer of many years standing, was able to arrange for a number of British applied linguists to visit the project. Some of these visitors not only wrote reports to the British Council but also talked about the CTP at conferences and devoted articles to it in the literature.

Johnson finds the CTP valuable because it "provokes thought concerning the notional syllabus" (1982, p.143). Roberts (1982, p.190) and Howatt (1984, p.288) anticipate that the project will serve as a test of the hypothesis that structure can be learnt without being specifically taught. For Brumfit, its value is based on the following considerations:

> a. it will have shown that a careful grass-roots experiment can be executed in a poor, third world education system in which experimentation is closely related to the activation of the teaching profession;
>
> b. it will have enabled us to obtain, in a non-idealised setting, valuable evidence about a major current model for language learning;
>
> c. it will have developed a set of materials which, with adjustments, can be used as a basis for fluency activities in any language teaching, regardless of whether or not the system is based on the underlying assumptions of the Bangalore project. (1984, p.235).

Davies regards the CTP as important to "current interests in Second Language Acquisition Studies" (1983, p.21). Allwright "valued the project very highly ... because of its bold attempt to investigate a hypothesis that has been much discussed ... that language learning is best seen ... as an unconscious process of acquisition through language use" (1981, p.1-2). Corder comments that

research into "interlanguage and second language acquisition ... gives a strong support to communicational approaches to language learning" (1982, p.1-2).

One clear thread runs through all of these views: the CTP had the potential to inform us about a major current model of language learning, that is, one which stresses unconscious processes and learning by using.

3.2.2 The Neglect of Evaluation

The interest, at least in terms of what has been written, has certainly been present in Britain. Academics in the United States, by contrast, although many of them are aware of the project, have not referred to it with enthusiasm, if atall, in journals and edited collections. Richards (1984) draws attention to it only to dismiss it; Long (1985), in an article about task-based language teaching, simply ignores it.

There are two possible reasons for this. First of all, there has been very little written about the project by Prabhu himself (although his recently published book [Prabhu 1987] may change the situation); in view of this, writers may have felt that there was little they could refer readers to. Certainly, an article of the author's which referred in passing to the CTP, drew the observation from the editor and a reviewer of TESOL Quarterly that an American audience would require a full introduction to the project as readers could not be expected to

have heard of it. Also, Greenwood's (1985) frustration with the CTP seems largely to derive from the lack of available published information.

The second reason may be that there is an unwillingness to find unsubstantiated opinions of very much interest. This has been frequently stated by both American and British colleagues in conversation, but occasionally in print, too. Crookes, for example, looking at the problems of judging the value of the CTP, says that "a point which even sympathetic commentators such as Brumfit have noted, is the lack of hard information about the success of the project" (1986, p.25). Richards selects the CTP as a classic example of "the need for rigorous evaluation procedures in planning methodological innovations" (1984, p.19). He inveighs against the inattention to evidence (also cited in chapter 4, section 4.2.1):

> Unfortunately, in the Prabhu study neither objectives nor evaluation was incorporated into the program design. This makes any serious consideration of his claims impossible. Carefully designed research takes neither more nor less time and effort to conduct than poorly designed research. (1984, p.20).

It is possible to quibble with Richards' conception of the value of objectives, but he is probably right to insist that if evaluation had been considered more comprehensively at the design stage, much more could have been learnt. The late attention to summative evaluation is referred to in chapter 4 as the 'point of entry problem', and the arguments in favour of early provision

for evaluation are presented there. Essentially, early provision enables baseline data to be collected, and the systematic monitoring of classroom practice to be effected; it also allows the various stakeholders to be made aware of the kinds of arrangements that would need to be made for certain types of information to be forthcoming, e.g. randomisation if internal validity is paramount. This is what is known in the evaluation literature as 'pre-evaluation' (Ross forthcoming), 'evaluability assessment' (Rutman and Mowbray 1983, p.37) or 'probing for a sensible charter' (Cronbach and Associates 1980, p.165). The collection of baseline data and the provision for systematic monitoring was quite absent from the developmental stages of the CTP and the attitudes to evaluation reported by Carroll in the RIE Newsletters and Bulletins.

It might be argued that from another perspective, evaluation entering the discussion at an early stage would serve to constrain development. That is to say, certain directions in program design might not be pursued because of a sense that returns would not show up in an evaluation. There may be a case for a program that wishes to evolve through trial and error to be given free rein until such time as the principles and methodology have settled somewhat and the program is prepared to declare itself in fairly precise terms.

However, at some stage, if external evaluation is

desired, it would be most useful if the program were to start afresh in new schools so that baseline data can be collected and implementation monitored. It happens that one of the visitors to the project in 1982, Douglas Barnes, made precisely these recommendations. The CTP, he said, "now seems to be ready to move to a more explicit account of itself" (1982, p.4). He advised setting the project up in new and different kinds of schools for at least four years; in addition, he recommended that detailed illuminative and summative evaluations be set in motion. The duration, the range of schools and the illuminative elements of the evaluation would promote extrapolation by potential users of the project, while the summative element would attempt to offer a more objective appraisal (for potential administrators, for instance).

Barnes clearly had in mind a rather large operation as he proposed that Dr. Prabhu should be assigned full-time to the project. This proposal may have been unfeasible, and the project could hardly have expanded without a full-time director. In the event, however, the project did continue with fresh schools - the 4 evaluated in chapter 4 - but all of Barnes' recommendations were ignored and never mentioned again. Even in 4 schools, they would have been worth following. It would appear that an opportunity for useful evaluation was missed at this juncture.

In chapter 5, full reference is made to the

ambiguity concerning the status of the project. Briefly, the description runs as follows: on the one hand, Prabhu states that the CTP does not see itself as an attempt at propagation, at usurping the dominant position of structural teaching in South India; he claims that CTP teachers were free to accept or reject or modify according to a personal sense of plausibility. On the other hand, in a public debate about the CTP (amplified in chapter 7), an extract from a CTP lesson provokes a reaction from some seminar participants to the effect that the lesson seemed just like a structural lesson; Prabhu ripostes that "in that case our burden of retraining is likely to be reduced" (1980a, p.50), thereby fuelling fears of a 'takeover bid'.

Davies (1983) and Brumfit (1984) both express doubts as to the true intentions of the project, and the author's own visit was also marked by conflicting impressions.

This uncertainty about what the CTP intended to achieve is reflected in the discussions on evaluation in the RIE Newsletters and Bulletins:

> what we expect at the end of this project is ... suggestions for evaluation criteria and methods, if the approach being tried out should ever be implemented on a large scale (RIE, 1979, Vol.1. No.1., p.3).

Evaluation here is clearly associated with some form of future large-scale implementation. Later, in a discussion about summative evaluation, the role of conventional

language tests is explored in connection with the CTP:

> Our view is that this type of evidence, while not necessary for us ... will probably have a role in convincing the language teacher of the effectiveness of our approach. We suggest, therefore, that until language teachers are confident of the value of the method, such evidence may be collected (RIE, 1980, Vol.1. No.4., p.29).

And, in the same discussion:

> 'General English' tests have some value. They will not be particularly difficult for our pupils, and although their overall validity as a test of language use is limited, such items are useful to persuade the 'hardened language teacher' of the effectiveness of the method (RIE, 1980, Vol.1. No.4., p.32).

From these two extracts, testing within a summative framework was clearly being considered from the perspective of persuading others of the effectiveness of the CTP, directly contradicting Prabhu's statement that "this is a searching exercise, not a selling one; an attempt at self-assurance, not at persuading others" (In RIE, 1979, Vol.1. No.2., p.21). In 1979 and 1980, summative evaluation appears to have been entertained, at least in part, as a means of 'persuasion' and large-scale implementation.

Even in 1982, and a month before Barnes made his recommendations, Corder visited the project and was under the impression that evaluation was to be linked to propagation:

> If this approach is to gain general recognition and eventually perhaps be adopted as an official teaching method then it is necessary to be able to show to the satisfaction of those who make the decisions that learning as or more relevant and effective takes place as a result of the

> teaching than occurs under the present generally practised 'structural' approach (1982, p.3).

It is unlikely that Corder would have made these comments had he not been satisfied that the CTP aspired to the role of an 'official teaching method'.

During Carroll's tenure in Madras (until 1981), the CTP received some kind of formative evaluation, later written up as a Ph.D.; this might have sufficed if 'self-assurance' were all that was of interest. However, summative evaluation for the purpose of winning over doubting teachers appears also to have been a matter of concern. In the end, a summative evaluation was sought (see chapter 4), but this was late in the day, and no reason has been offered for the neglect of Barnes' proposals. The only conclusion that presents itself is that the inconsistencies concerning the intentions of the project are reflected in the inconsistencies evident in the attitudes towards evaluation.

3.2.3 The lack of pupil-pupil interaction

One of the major criticisms the CTP has faced relates to the almost total lack of pupil-pupil interaction. Barnes, for instance, noted that "most of the teaching in the CTP lessons consisted of teacher-class exchanges" (1982, p.2), and he suggested the "use of work in pairs and small groups, in which all utterances are not channelled through the teacher" (1982, p.3).

Howatt anticipates that "some of its [the CTP's] characteristics will cause comment if not controversy, in particular, the low priority it attaches to social communication" (1984, p.288). Brumfit, too, feels bound to comment on this matter: "it does need to be said that by no means all the students participate overtly ... it seems that group work would increase the chances of active participation" (1984, p.237).

Aware of these objections, Davies takes a somewhat different view. He points out that "criticisms of the CTP that it cannot be communicative or indeed communicational because it does not involve pupil-pupil interaction are beside the point since this is the way the CTP is set up" (1983, p.6). This seems fair comment as it is hardly the business of an innovative project to follow prescribed notions about methodology. However, Davies, in a talk to M.Sc. students at Edinburgh University in 1983, drew attention to the fact that only a relatively small number of students responded to the teacher. The author's own observation of a number of lessons in 1984 confirmed this. Brumfit's argument that group work might therefore encourage more pupils to participate overtly does have substance.

Indeed the CTP had no objection atall in the early years to group work. On the contrary, towards the end of the first year of the project, it was viewed with considerable favour. A lesson on December 3rd 1980 prompted the CTP team to take up group work forthwith;

few students (only 22) attended the lesson because that day was a local holiday, and the wider participation by students "brought home to the group the fact that one of the major problems of teaching on this project has been that of handling a relatively large class"; the CTP team therefore decided that "an attempt should now be made to move on from lock-step procedures to group tasks" reported in RIE Newsletter Vol.1. No.3., pp.24-25). In the same Newsletter (p.30), it is thought that group work "would provide increasing opportunities for meaningful interaction among the students themselves, in place of the exclusive practice of dialogue between teacher and class".

By the 29th January, the CTP team found "the degree of commitment shown by all the students to the idea of working in small groups" encouraging (p.52). At a meeting on 19th February, it was concluded that:

> the transition from whole-class, lockstep procedures to group work (...) has been more successful than it was in earlier attempts. Not only do the students now seem to enjoy team work (...) but the teams generally function better as teams - i.e. with members more willing to participate/contribute, instead of seeking to hide behind a leader. (p.57).

At the same meeting, it was also noted that "a big problem in teaching by lockstep procedures" was "the varying pace at which different students were able to work" (p.58). At this stage in the CTP's development, group work was found to be accepted by the students, helpful in overcoming large class-size and encouraging

wider participation, and possibly useful in dealing with different ability levels.

Others working in India had long since come to similar conclusions. In a single issue of Teaching English: a magazine devoted to the teaching of English in India in 1955, the following observations were made by different authors: Forrester, (who had much to do with the implementation of the structure-based sylllabus in Madras), recommended group work as a way of involving large classes and a wide variation in ability; like the CTP team, she could claim that it had been useful: "the writer has used group work successfully" (1955, p.8). Hensman stated that "even the most backward in the class can share learning very freely with others because they get more chances to speak" (1955, p.21). Billows adds his voice: "to overcome the difficulties of teaching over-large classes teachers should be encouraged to break up their classes into groups" (1955, p.25).

By the time of the annual review seminar of 1980, the practical benefits of group work that had been perceived were being offset by apparent theoretical doubts. Would group work result in the loss of struggle to say something? Was there a problem in that some learners were more advanced than others? At this time, these questions weres merely doubts: "we have not done enough thinking about this; there might be a way out" (RIE Bulletin, 1980, Vol.4. No.1., p.163).

The next mention of group work comes in 1982, by which time there are no further doubts; group work is out. Prabhu maintains that it would lead to pidginisation, that it would result in students using L1, and that it would be too severe a break with tradition (1982, p.6). The objections do not include the possible loss of struggle to say something, but refer to 3 elements not noticed in the earlier practice with group work nor in the discussions reported in the Newsletters (although the worry about some learners being more advanced than others could have led to the view about pidginisation).

This is perplexing enough, but is difficult to know what to make of Prabhu's more recent statements regarding group work. He says:

> the avoidance of group work in a more organised form was, at the beginning of the project, due to a wish to confine pedagogic exploration to the project's major principle (...) but more positive reasons for excluding it came to be perceived in the course of the project. (Prabhu 1987, p.81).

This gives the impression that there had always been some objection to group work and that it had always been avoided.

These discrepancies are puzzling and are merely noted here. The project's central tenet that form is best acquired through a focus on meaning does not require the adoption or rejection of groupwork. Clearly, the technique may have practical benefits and theoretical positions may be proposed, but they are, as Davies (1983, p.6) observes "beside the point". Nevertheless, the issue

has been thought important by some commentators and attention is duly accorded to it in this critique, in which discrepancies in CTP attitudes to groupwork are documented.

3.2.4 Coverage

Davies comments that not only is the lack of learner-learner interaction a constraint in CTP teaching but also the "reliance on referential language and not atall on affective language" (1983, p.6). That is to say, the 'coverage' of language that the CTP apparently makes provision for is restricted. This has been one of the principal criticisms of the project.

Barnes complains that the "range of rhetorical functions was highly restricted, such functions as persuasion, argument, cross-questioning etc. not occurring" (1982, p.2). This is precisely the point that Johnson makes in the review seminar in Bangalore in 1980: "it is possible that at the end of your course students will never have practised things like apologising, expressing surprise, etc." (In RIE Bulletin, 1980, Vol.4. No.1., p.152). Prabhu replies: "there is no real danger because learners will be learning to learn; if they listen to apologies they will pick it [the relevant language] up" (p.153).

Prabhu elaborates on this notion and his views are worth quoting in full:

> it seems very unlikely that communicational teaching over a span of one or more years will fail to take in any of the basic elements of the language; or, putting it conversely, any elements which are never called for in such teaching over such a span are unlikely to be very 'basic' ... if learners learn some parts of language structure without there being any deliberate/specific attempt to teach those parts, one can say that they have learnt to learn language-structure in the process of having to use it and will therefore learn the rest of it when they find the need for it. (Prabhu, in RIE Newsletter, 1980, Vol.1. No.4., p.18).

This is worth quoting in full mainly because it helps to make clear that the arguments Prabhu advances are unfalsifiable. Prabhu states that all relevant language will be covered in the course of a year or so. It is impossible to know this since he gives no specification of what range of language this might be and no provision was made to record classroom language systematically. So now we can never know (at least in relation to the CTP). He then proceeds to argue that if learners learn some language without being deliberately taught, they will learn the rest when the need arises. But of course, as the 'how far' lesson extract and related discussion clearly indicate (see chapter 7, section 7.1 for full comments), it is impossible to know whether language was taught 'deliberately', and even if this were known, it would not be possible to say whether the deliberate teaching or something else was the cause of the learning. But even if all of this were susceptible to some form of proof, an inductive leap is required (viz. the 'rest' of language structure will be learnt).

It would be excessively prescriptive to discourage scholars from indulging in speculation, but it is equally important to acknowledge that an unfalsifiable argument is hardly a powerful one. All that this means is that the charge that the CTP does not ensure coverage has not been satisfactorily answered.

Bound up with the question of coverage is the reliance on reasoning that has troubled some observers. The reliance on reasoning, that is to say, has a bearing on the lack of coverage. If tasks involving affective abilities were included in the CTP, it is thought that this would increase the rhetorical range.

Brumfit lists as a major criticism of the CTP that it "relies too much on student reasoning" (1984, p.236). Prabhu (1982, p.6) puts forward 4 arguments for this reliance: (i) that learners feel more security with right/wrong answers, (ii) that it encourages guessing (seen as a desirable strategy in the CTP), (iii) open-ended questions make greater demands on learners' language than is deemed appropriate at an early stage, and (iv) that in any case English is the language of rationality and not of emotion for Indians.

No formal survey has to my knowledge been carried out on this last assertion, but those Indians asked casually by the author have been mystified by it. If this point is really to be considered seriously, it needs to be substantially clarified.

As for the matter of open- versus closed-ended questions, Brumfit points out that it is "not strictly the same as the rational-emotional distinction" (1984, p.237). Moreover, he demurs at Prabhu's assumption that "learners will be motivated by an essentially intellectual curiosity" (1984, p.237). Finally, he says that Prabhu probably does not follow his own precept, as some of the problems require students to respond imaginatively (1984, p.237).

Prabhu's point that greater demands are made on learners' language than is appropriate at early stages of learning meets with some sympathy from Barnes (1982). However, Barnes stresses that

> there seems to be every reason for the CTP materials to include a progressively greater proportion of open-ended activities as the course proceeds ... in later years these activities could be arranged to include a wider range of rhetorical functions. (1982, p.3).

What Barnes regards as important is not only that coverage needs to be considered further, but that learners should also have the opportunity to produce language in a less constrained way at some stage. None of the commentators cited here has argued that production should be forced, but merely allowed. (Chapter 7, section 7.5.4.4 explores this notion further).

3.2.5 The Possibility of a Hidden Syllabus

The CTP ostensibly has no language syllabus, but Johnson (1982) has argued that it may, in fact, have

arrived at some form of linguistic syllabus, involving such elements as 'distances' and 'directions'. More precisely, Johnson contends that since Prabhu's procedural syllabus grades tasks conceptually, it is open to the charge that "it is in fact a covert semantico-grammatical syllabus" (Johnson 1982, p.140).

Both the fact that CTP tasks are described in such terms as 'directions' (i.e. the cardinal points on a compass) and the fact that the pre-task introduces such language as may be necessary to tackle the task increase the similarity. Johnson focuses on the pre-task, suggesting that this is the stage in a lesson when pre-teaching of concepts and related language might become linguistic pre-specification (1982, p.141). This seems a valuable point to make. If teachers know the task is about the distances between cities, it is quite feasible that they will all pre-teach 'how far...?'. If each teacher teaches the same lesson to different classes, it is conceivable that the pre-teaching will be quite consciously based on a certain amount of linguistic pre-specification.

Commenting on Johnson's argument, Brumfit objects that it is not necessarily the case that the procedural syllabus is a covert semantico-grammatical syllabus (1984, p.239). He points out that first of all, Prabhu's concepts are not stated specifically; secondly, that while some of the realisations of the teacher's talk will coincide with certain semantico-grammatical categories,

they could also be realised in a wide range of grammatical structures. That is to say, the procedural syllabus is not systematic in semantic or grammatical terms; it is not sequenced in these terms.

Perhaps Johnson's point is better made without the reference to a syllabus type. He does seem to have a case in saying that linguistic pre-specification is a possibility in the CTP.

Davies, in fact, makes a very similar point. However, where Johnson argues that the specification is a matter of design, Davies suspects that it may be arrived at by default. His concern is that typical teachers in South Indian schools are lacking in proficiency in English, and that this could have a bearing on the way that the CTP operates in the classroom. In his observation of CTP lessons, he notes the "occurrence of holophrases" (1983, p.10), and sees problems in the teacher-class interaction:

> The difficulty for the teacher is to operate a negotiating channel with the pupils: otherwise the CTP 'syllabus' can be as rigid (and uncommunicative) as a structural syllabus. Thus there is in cases of inadequate proficiency a temptation to stick close to a 'script' (1983 p.12-13).

If teachers with low English proficiency do stick to a script, then the implications for the external validity of the project are serious. (The question of how well regular teachers are able to implement the CTP is looked at in some detail in chapter 5).

3.2.6 Research carried out on the CTP

Apart from the research reported in the present thesis, a number of studies have been carried out on the CTP. 4 M.A. students at the University of Lancaster made some analyses under the supervision of Richard Allwright (who had visited the project). Also, one M.Sc. student at the University of Edinburgh carried out research under the supervision of Alan Davies (who had also visited the project). This body of work will now be considered, starting with the Lancaster students.

The 4 Lancaster students are Gilpin, Collingham, Mizon and Kumaruvadivelu. Their studies were written in 1981, and were based on audiorecordings of the pre-task phase of 4 CTP lessons in Bangalore and audiorecordings of 4 parallel lessons taught in a primary school in Dalston, Cumbria (i.e. L1 teachers to L1 learners). Two teachers in the Dalton school were presented with exactly the same lesson materials as were used by the CTP teachers and told to accomplish the pre-task phase in any way they saw fit. The recordings were then transcribed and subjected to various analyses by the 4 M.A. students. The purpose of the studies was to compare the teaching strategies and the laguage used in the different circumstances.

Gilpin analysed the 4 comparison lessons using a modified version of Bellack's (1966) framework. She found that there were a higher number of general solicits from

Bangalore teachers than from the teachers at Dalston, and that Bangalore teachers tended to accept multiple responses; the Dalston teachers made personal solicits. The analysis also revealed that Bangalore lessons had a higher frequency of 'react' moves than Dalston.

In addition, cycle analysis disclosed that there were far more extended cycles in Bangalore than in Dalston. She suggests that a reason for this discrepancy may reside in different lesson structures. In Dalston, the processes required to solve problems are established by the teachers first, whereas in Bangalore the processes emerge from modelled examples during the lesson. Establishing the processes first means that reference can be made back to them in the form of simpler solicits, generating frequent new cycles. Gilpin proposes that since the RIE Newsletters state that basic problem-solving strategies - even in L1 - are undeveloped in the CTP, surely greater overt guidance would be worth considering.

Mizon looked at only 1 lesson from Bangalore and the parallel lesson from Dalston. She carried out a lexical analysis, a syntactical analysis and a contextual analysis of reformulation and repetition in teacher talk.

Her study revealed that a greater variety of verb structures are used in the Dalston classroom, and that although the use of 'one main verb' in the 'present simple' is the most common verbal structure used in both classrooms, the frequency is relatively higher in the

Bangalore classroom.

Interrogatives were more often used than declaratives in Bangalore, and there is a relatively greater occurrence of repetition and reformulation strategies in Bangalore, too. Her study was purely descriptive and she felt that it could have little generalisability value; therefore no attempt was made to "establish a relationship" with the "project's theoretical framework" (Mizon 1981, p.36).

Kumaruvadivelu, also analysing 1 lesson from each location, considered 'turn-taking'. He investigated turn-getting and turn-giving strategies. Among his findings was that mean length of turn in Dalston was far greater than in Bangalore, but that there was a greater variety of sequences of turns in Bangalore where the teacher appears to exhibit more flexibility in varying types of solicit according to types of elicitation. He makes a few suggestions for involving the less competent learners.

Collingham, also analysing 1 lesson from each location, uses Mehan's (1979) framework of classroom interaction. After initially dividing the pre-task phase into topic-related sets, she then proceeds to distinguish basic from extended sequences, and found that the Bangalore lesson incorporated proportionately more extended sequences than Dalston, and less basic sequences.

Taking a closer look at the extended sequences, she

divided the teachers' negative evaluations into a number of strategies suggested by Mehan (1979). The main differences were in the frequency of prompting (more in Bangalore), repetition/reformulation (more in Bangalore) and process questions (more in Dalston).

(The descriptions and examples of 'basic' and 'extended' sequences, and of the various strategies are given in Appendix 1).

One point to emerge from both Gilpin's and Collingham's studies was that there are more extended sequences in the Bangalore lessons than in Dalston. Given that more lessons were available to the author, Collingham's study was replicated on the larger data set, and this study is reported in Appendix 1. Incidence of the various strategies used varied somewhat in the larger study, but Colllingham's finding in relation to basic and extended sequences was corroborated. There are many possible interpretations of this: perhaps the Dalston students found the problems easy; perhaps the Bangalore teachers were deliberately asking questions that would require more of a struggle in order to bring about a preoccupation with the task (although, if that were the case, why not select a more difficult problem?); it may also be that the Bangalore teachers could benefit from greater attention to the structuring of content lessons.

But what can be learned from the Lancaster studies? Allwright (1981) recommended that the project be given assistance with respect to classroom research techniques,

and noted that this would be accomplished through analyses of classroom recordings by postgraduate students at Lancaster. Use could possibly have been made of the studies as a contribution to formative evaluation (as has been noted, the studies did make suggestions), but there has never been any indication that issues raised by the studies (e.g. the structuring of the lessons) were fed back into the project.

With regard to their contribution to a summative evaluation, the data base is far too small to permit more than the most tenuous of interpretations. This is candidly acknowledged by all 4 researchers. Also, the Dalston school is hardly comparable with the Bangalore schools.

Apart from a brief reference by Brumfit (1984), these Lancaster studies have not been referred to in the literature except by Prabhu. Responding to Greenwood's (1985) suspicion that the teaching of structure and vocabulary was probably done consciously on the CTP, Prabhu says:

> I can perhaps refer him to an investigation made by four post-graduate students of applied linguistics at Lancaster University (...) in 1981, which consisted of getting a subject-teacher at a British school to do, with a class of British children, four of the tasks used on the project in India, and comparing full transcripts of the resulting lessons with those of the corresponding lessons in India, to see if there was evidence in the latter of covert teaching of language items. (1985, p.77).

If these studies are being offered as evidence that

covert language teaching did not take place, then this is a far greater claim than the data can support, and far more than any of the 4 M.A. theses has claimed.

The study carried out by Saraswathi (1984) at Edinburgh University will now be considered briefly.

Saraswathi's study had 2 major strands: "(a) analysis of the process by which the learner is expected to achieve grammar construction" and "(b) analysis of the product, i.e. the extent of grammar construction achieved at the end of two years of the CTP - both receptive and productive" (1984, p.32). Her data are a set of task outlines.

She finds in connection with the first area of inquiry that lessons in a second cycle of lessons would require "higher cognition" (1984, p.61) than those in a first cycle. Basically the tasks become more difficult as they are re-introduced. With regard to the second area of inquiry, she finds that the linguistic competence that the tasks demand in later stages of a course is of a higher order than that expected at early stages.

It is worth bearing in mind that the data were only lesson outlines. How far the outlines reflect the lesson input is not known. Furthermore, to what extent learners actually achieved the linguistic competence the lesson outlines are thought to demand is also unknown. Aware of these limitations, Saraswathi appeals to her retrospective impressions:

> the researcher was an observer (or a teacher) of every one of the lessons and found that the learners were with the teacher; they did understand what was happening and engaged in the problem-solving activities with great enthusiasm. Further the feedback from each day's correction of tasks suggests that in general more than half the class answered more than half the task correctly. (Saraswathi 1984, p.74-75).

This is of interest as an opinion, but as far as the study is concerned, the findings are necessarily limited.

Saraswathi's study, it is worth mentioning, is a useful source of reference as it provides a coherent account of the CTP, and is, after all, written by an 'insider'.

3.3 Summary

This chapter has attempted to accomplish two aims. First, to provide an account of the CTP such that its evolution and its stated principles and methodology are clear. Second, to bring together all of the literature that has been spawned by the project in disparate sources, and to try to analyse critically their contribution to our understanding of the CTP.

An impression that emerges strongly from a reading of the literature is that the CTP has provoked a great deal of controversy, and this chapter has aimed to document the major areas that have fuelled the arguments for and against. These areas are (i) the apparent neglect of hard evidence in the form of systematic evaluation, (ii) the lack of learner-learner interaction, (iii) the reliance on referential language and (iv) the possibility

of a hidden syllabus.

The great interest and excitement that the project has stirred is also reported (section 3.2.1), as are the small scale studies carried out by M.A./M.Sc. students at the Universities of Lancaster and Edinburgh.

It was noted that most of the commentators on the CTP saw the project as primarily interesting for the potential it had to inform about a major current model of language learning - one that stresses unconscious processes. Well conducted experimental studies might conceivably assist in the development of such theoretical models, but the project lacked control (as chapter 4 makes clear) and from the very beginning of the present evaluation, it was clear that it was beyond the scope of an evaluation to provide proof for a learning theory.

Instead, the aim of chapters 4, 5, 6, and 7 will be to provide information about the CTP that will facilitate extrapolation to other circumstances.

CHAPTER 4

A COMPARISON OF CTP AND STRUCTURE-BASED TEACHING

4 A product comparison of the CTP and the prevailing structure based approach in S. India

4.1 Introduction

This chapter describes a product evaluation of the CTP. This segment of the evaluation was launched in February 1983 when Dr. Prabhu of the British Council in Madras invited Alan Davies of the University of Edinburgh to advise on it, and to attend a CTP seminar in Madras. Davies' subsequent report to the British Council (Davies, 1983) set out a range of possible designs for the evaluation, and recommended that a research student (the author) should prepare test instruments under Davies' guidance during the autumn of 1983, prior to administering the tests in situ during the first three months of 1984.

During the autumn of 1983, Prabhu was asked to specify precisely what his purposes were in seeking an evaluation so that the design of the inquiry could meet his requirements. His stated purpose in seeking the evaluation was:

> To assess, through appropriate tests, whether there is any demonstrable difference in terms of attainment in English between classes of children who have been taught on the CT Project and their peers who have received normal instruction in the respective schools. (Prabhu, 1983, personal communication).

This chapter, then, deals with the testing element of the evaluation.

4.2 Constraints

4.2.1 The Point of Entry Problem

The CTP began in 1979. The invitation to evaluate was made in 1983 for the tests to be administered in 1984. It was seen by the CTP team that the evaluation would signal the end of the project, or at least its prevailing phase. In other words, evaluation was perceived as a means of wrapping up the project and obtaining some measure of its success. This is what Scriven (1981) has called the 'point of entry' problem, which is a problem of when an evaluator should be brought in on a project.

For Scriven, the main difficulties consequent upon evaluation being sought too late in the project's life are the impossibility of (i) obtaining the baseline data, (ii) establishing experimental control, and (iii) determining gains or causation. Quite apart from these purely experimental design considerations, late entry means that the evaluator will not have made adequate provision for recording what transpired between the inception of a project and its termination. This is especially important since true experimental design is not usually feasible in educational research, and therefore some way of explaining results is needed which is not causative. So, for example, it would be helpful to audio-record a large sample of lessons over time.

Analysis of these lessons would provide a valuable guide as to the relevant variables, and would aid interpretation of test results.

Stufflebeam puts the point of entry problem succinctly:

> The most usual point of entry problems are those that arise because an evaluation was requested too late. For example, a project director suddenly decides near the end of a project cycle that an evaluation is needed ... In such a case, the evaluator is asked to work a miracle by somehow obtaining data whose opportunity for collection has long since passed. (1985, p.124).

This is such a common problem that Stufflebeam takes the view that, in the short term, evaluators must simply decide "how to make the best of a bad situation" (1985, p.125).

It is quite clear that the CTP suffers from the point of entry problem and it has not escaped censure for this. Richards chides Prabhu for failing to consider evaluation in the developmental stages of the CTP:

> Unfortunately, in the Prabhu study neither objectives nor evaluation was incorporated into the program design. This makes any serious consideration of his claims impossible. Carefully designed research takes neither more nor less time and effort to conduct than poorly designed research. (1984, p.20).

Insofar as this criticism relates to the need for a program to take evaluation into account at the design stage, it is well-founded.

Apart from the fact that late entry implies an absence of baseline data and an adequate description and analysis of process, in the case of the CTP it also meant

that tests were initially produced in a vacuum, i.e. without the opportunity for piloting. Thus, once the types of tests had been decided, level could only be guessed at. Three levels were devised for each test, but test construction had to continue in India under pressing constraints of time, in order to tailor the level to the learners' abilities.

4.3 Procedure

4.3.1 The Schools

Four schools were involved in CTP teaching at the time of my visit in January 1984. They all agreed to take part in the evaluation:

(a) Bangalore (St. Antony's Kannada Upgraded Primary School, Jayanagar, T-Block, Bangalore 560 041)

(b) Cuddalore (Sacred Heart's Tamil-Medium Primary School, Cuddalore, Tamil Nadu)

(c) T.Nagar (Sri Sharada Vidyalaya Middle School, T. Nagar, Madras 600 017)

(d) Tiruvottiyur (Vellayan Chettiar Higher Secondary School, Tiruvottiyur, Madras 600 019)

In each school, two classes were to be tested: the CTP class and a peer-group class who had been taught by a structure-based approach (i.e. the regular English teaching approach used in South India).

Three of the schools are mission schools (Tiruvottiyur is the exception), but they are in no way

elitist. The CTP had previously been tried out in a series of Corporation and Government schools, but it was found that discipline and administration were incompatible with the basic requirements of a developing experiment. For example, in a Bangalore school in 1981, 55 pupils began the project classes but because of a high drop out-rate from the school and constant internal reshufflings, by the end of the year only 18 of the original group remained. Moreover, they had completed only 90 instead of the normal 130 lessons in a school year.

4.3.2 The Pupils

In this section, the pupils in both experimental and control groups at each of the 4 schools are described.

Bangalore:

(i) Experimental group: A standard VI class of 11/12 year old boys and girls. The class had been taught English through the CTP from the beginner's level for two academic years (1982-1983 and 1983-1984), about 250 lessons in all, of 40 minutes duration each.

(ii) Control: The other standard VI class (there were only two). However, for want of space at the school, this class functioned at another site about a mile away. This class, too, consists of 11/12 year old boys and girls. They had been taught English for the same length of time (viz. two years) by the structure-based methods prevalent

in the state.

Cuddalore:

(i) Experimental group: A standard IV class of 9/10 year old girls. The class had been taught through the CTP from the beginner's level for two academic years (1982-1983 and 1983-1984), about 275 lessons in all, each of 40 minutes' duration.

(ii) Control: One of the other two standard IV classes at the same school, consisting of 9/10 year old girls. They had been taught English for the same length of time (2 years) by the structure-based methods prevalent in the state.

T.Nagar:

(i) Experimental group: A standard V class of 10/11 year old boys and girls, who had been taught English through the CTP from the beginner's level for three academic years (1981-1982, 1982-1983 and 1983-1984), about 380 lessons in all, each of 40 minutes' duration.

(ii) Control: One of the other standard V classes in the same school, consisting of 10/11 year old boys and girls. This class had been taught English for the same length of time (3 years) by the structure-based methods prevalent in the state.

Tiruvottiyur:

(i) Experimental group: A standard VI class of 11/12 year old boys who had been taught through the CTP for one academic year (1983-1984, about 120 lessons of 40 minutes duration each) in their fourth year of English. They had been taught earlier for three academic years by the structure-based methods prevalent in the state.

(ii) Control group: Another standard VI class of 11/12 year old boys in the same school, who had been taught for four years by the structure-based methods prevalent in the state, though during the fourth year, this was done consciously.

The experimental groups described above are four of the eight classes taught (for a year or longer at different schools) through the CTP. The project team had to stop teaching the other four classes at various times for reasons such as the class having reached the stage of taking a public examination or the school having to reorganise its classes as a result of a high drop-out rate. (Three of these classes which took public examinations did slightly, though not significantly, better than control students, thus establishing that at least the project was not disadvantaging experimental students).

4.3.3 The Teachers

In this section, the teachers who taught either experimental or control groups in the 4 schools are described.

Bangalore

(i) The experimental group was taught largely by a a lecturer at the B.E.S. College of Education in Bangalore, whose own M.Ed training included methods of teaching English. This teacher also had an M.A. in English Literature. The rest of the teaching (about 25%) was done by a regular teacher at the school, who had a teaching diploma.

(ii) The control group was taught by a regular teacher at the school, who had a teaching diploma.

Cuddalore

(i) The experimental group was taught largely by a tutor at the English Language Teaching Centre at Cuddalore, who had M.A. and M.Ed. degrees. A regular teacher at the school, who had a teaching diploma, did the rest of the teaching (about 15%).

(ii) The control group was taught by a regular teacher at the school, who had a teaching diploma.

T.Nagar

(i) The experimental group was taught in the first two years largely by a member of the original project team whose qualifications include a Ph.D. in Linguistics from the University of Reading, but who had no earlier experience of teaching at the school level. In the third year, a large part of the teaching was done by another member of the project team, whose qualifications include

a master's degree, a teacher training diploma, and an M.A. in Applied Linguistics from the University of Lancaster. This teacher had had 12 years of ELT experience before the project began. The rest of the teaching, about 30% in the first two years and 25% in the third year was done by a regular teacher at the school, who had a teacher training qualification.

(ii) The control group was taught by a regular teacher at the school who had a teaching diploma.

Tiruvottiyur

(i) The experimental group was taught by a regular teacher at the school, who had a B.A. in history.

(ii) The control group was taught by a regular teacher at the school (whose qualifications are unknown).

All the teachers who taught experimental groups (other than the two project-team members at T. Nagar) were given professional assistance in teaching to CTP requirements in the form of (i) 3 demonstration lessons at the beginning, taught by a member of the project team and discussed in some detail, (ii) the collection of the tasks developed on earlier project teaching, to choose from or adapt in their own teaching and (iii) periodical visits to the class by one or another member of the project team who observed a lesson, taught a lesson or made some suggestions.

4.3.4 The Tests

4.3.4.1 Alternative Testing Strategies

In his report to the British Council, Davies (1983) proposed 5 alternative testing strategies: (i) Test the CTP group on an achievement test, (ii) test CTP and control groups on both CTP and structure-based tests, (iii) test CTP and control groups on a 'neutral' test, (iv) test only the CTP group on a 'neutral' test, and (v) test CTP and control groups on parallel forms of a test.

4.3.4.2 Program-Specific and Program-Neutral Tests

Once Prabhu had specified the purposes of the evaluation, proposals (i) and (iv) were immediately dropped. However, there seemed no way of testing the comparison programs fairly. The problem of program-fair appraisal, as has been argued extensively in Chapter 2, section 2.2, has so far proved insoluble. The options employed by previous researchers are: standardised tests, specific tests for each program, common/unique objectives and an appeal to consensus. Recalling the argument put forward in chapter 2, standardised tests were rejected outright for the present study because of their insensitivity to features of specific programs; the objectives model was also rejected because the comparison programs could only be said to have broad, long-term goals in common, but in the short run, different kinds of achievement were expected; an appeal to consensus would

mean an arbitrary decision about how language is best learned, and was therefore inappropriate to a comparative inquiry; there remained only the option of administering specific tests for each of the comparison programs which means that one program must be so superior to another that learners perform significantly better on both tests, an unrealistic demand.

One further possibility raised tentatively by Davies (1983, p.18-19) and Brumfit (1984, p.238) was that comparison groups could be tested on the state public examinations since the CTP approach claims to be effective in teaching the structure of the language. For this to be a viable alternative, it would have to be believed that the nature of the grammars learned under CTP and structure-based methods can be equated at early stages of learning, and this seems problematic.

All pupils taking part in the evaluation were at fairly elementary stages in their language studies. Structural pupils are intended to achieve mastery level over a limited set of structures prescribed by the syllabus for each year. Communicational pupils are not intended to achieve mastery level until, presumably, nature has taken its course (and there is no claim that this happens at the elementary level). A conventional grammar test (such as the Karnataka and Tamil Nadu public examinations) measures attainment or lack of attainment of mastery. That is to say, it measures at the level of a fully-formed competence a prescribed quota of structures.

The CTP makes no claim of uniform attainment, either as to which structures will be assimilated or about what stage of development learners will have progressed to. Therefore, at an elementary level, to compare both groups on an elementary grammar test would be perverse. It would be to count the CTP chickens before they have hatched.

If, on the other hand, the evaluation were taking place with advanced level students, then the notion of mastery would be more applicable to both groups, because by that stage, pay-off in such terms could plausibly be demanded. Otherwise 'incubation' would have to be dismissed as a luxury schools cannot afford. Given, however, that the CTP involved only elementary levels, the public exam option could not be justified.

It was hoped that a way out of the testing quandary could be found by catering both for program-specific features (Davies' second proposal) and features that may be said to be program-neutral (the third proposal). Thus, tests were devised that were specific to each program (program-specific achievement measures) and in addition, tests which did not take their cue from either program (program-neutral proficiency measures). (The tests and the rationale for them within this strategy are discussed in the following sections, 4.3.4.3 and 4.3.4.4).

4.3.4.3 Description of the Tests

The tests intended to measure achievement separately

for experimental and control groups were (1) a structure-based test and (5) a CTP task-based test. Tests intended to measure proficiency and to be syllabus-neutral were (2) contextualised grammar, (3) dictation and (4) listening/reading comprehension. (The full tests and marking schemes are presented in Appendix 2).

(1) Structure-based test

The structure-based test drew items from the lists of structures in the Tamil Nadu and Karnataka state syllabuses which the 4 schools adhered to. The test consists of a series of multiple-choice items.

Example:

We _____ going to school today. It's Sunday.

a. aren't b. not c. isn't d. don't

(2) Contextualised grammar test

This comprised a number of items where the testee was required to fill in the blank with one word.

Example:

Through the window I can see my father. He can't see me because he _____ looking at the road. He is going to the market.

In these items, often more than one answer was possible, so a checklist of correct and acceptable responses was drawn up (see Appendix 2).

(3) Dictation test

A short passage was dictated in the following way:

(a) reading of whole passage at conversational speed; e.g.

I have two brothers and three sisters. We all go to the same school. Sometimes we take the bus. Today we are going by bus. After school we will walk home.

(b) one reading only of each segment at conversational speed;

(c) final reading of whole passage at converstional speed.

(4) Listening/reading comprehension

This required testees to read, for example, a hotel advertisement and to write answers to spoken questions. It demanded a great deal of inference; e.g.

> Hotel Ashok: One room only Rs. 150 a day! Bring your family! In our grounds you can enjoy cricket, football, and kabaddi. We have a good restaurant. English and Indian meals. Film show every night at 8.p.m. Write to Hotel Ashok, 74 Gandhi Street, Delhi. Tel: 883921.
>
> Listen carefully to the questions. You will hear each question twice. Answer the questions using the information from the advertisement.
>
> E.g. spoken question: Where is the hotel?

(5) CTP Task-based test

This test was a representative sample of the tasks used in the CTP classrooms, as recorded by the RIE Newsletters, Bulletins and Lesson Reports. For example, solving problems related to a timetable and to a calendar.

4.3.4.4 Rationale for the Tests

The two achievement tests, (tests 1 and 5), were designed to reflect the teaching that had taken place in the comparison classes. Results on these tests could be expected to indicate whether what was learnt in the structural groups greatly diverged from what was learnt in the CTP classes.

Turning now to the proficiency tests, the aim was to select test-types which would be in some sense syllabus-neutral. Clearly, this is a vague concept, so it was important to have an a priori rationale to the effect that both groups could reasonably compete on each of the test-types selected.

Relevant to test 2 is Krashen and Terrell's observation on tests of contextualised grammar that:

> While it is possible that the student will understand the meaning and fill in the blank on the basis of acquired knowledge, it is also possible that the student will simply figure out the morphological pattern ... without even understanding the text. (1983, p.167).

If this is true, then both CTP and structural groups could be reasonably fairly compared on a test of this nature.

The justification for a dictation test 3 in this context rests on the theory proposed by Oller (1979 and elsewhere) that it measures a learner's 'grammar of expectancy'. He maintains that if the segments are too long to be memorised and regurgitated, they must be reconstituted by drawing on a grammar of expectancy.

Performance is therefore more or less successful depending on the sophistication of the learner's grammatical competence. Dictation can also be regarded as a sentence-bound test, thereby measuring structural awareness. In either case, dictation does seem to be a test suitable to both experimental and control groups.

The listening/reading comprehension test (4) is a test of receptive ability to use language. Its function here is to determine how far what is learnt in structural and CTP classes can be mobilised.

4.3.5 Hypotheses

The rationale for the tests can be restated as three hypotheses to be confirmed or disconfirmed by the results of the tests. These may be formulated as null hypotheses:

1. There is no difference between the language abilities arising from form-focused teaching and those resulting from meaning-focused teaching.

2. Acquisition of non-syllabus-based structure is not best achieved without focus on form.

3. Structure acquired without focus on form is no more readily available for deployment than structure learned with focus on form.

The null hypotheses could be rejected by the following test outcomes:

For the first null hypothesis to be rejected, there would have to be a significant difference in the performance of the comparison groups on the achievement tests 1 and 5. Either the experimental or the control group would need to perform significantly better on the CTP test, the structure test or both. If there were little difference on either test, it would suggest little difference in the language abilities arising from the different teaching approaches, and the null hypothesis would be accepted.

For the second hypothesis to be rejected, experimental classes would have to do significantly better than control classes on the proficiency tests of contextualised grammar (2) and dictation (3). If this second hypothesis were rejected, the central CTP hypothesis would be borne out, i.e. that form is best acquired through a focus on meaning. If there were no significant differences, the null hypothesis would stand.

For the third hypothesis to be rejected, CTP classes would have to score significantly higher than control classes on the proficiency test of listening/reading comprehension (4). No significant difference would mean acceptance of the null hypothesis.

4.4 Experimental Design

Before the results are considered, it is first of all necessary to examine the degree of control that the product evaluation achieved, as this would influence

interpretation.

In the present inquiry, CTP classes were regarded as experimental and the structural classes as control. A 'true' experiment would require students to be randomly assigned to experimental and control groups in order to ensure that the groups were initially equivalent. It is then a matter of choice as to whether or not to use a pretest-posttest design or a posttest-only design to measure gains. However, as has already been noted, the evaluation was only sought towards the end of the CTP's life, so that there was no opportunity to assign pupils randomly to programs, nor to measure their initial equivalence. Therefore, it was necessary to opt for a less rigorous design involving intact classes.

Precisely because full experimental control was lacking, it is especially important that the specific variables the chosen design failed to control be made explicit. To this end, two checklists of factors affecting internal and external validity were drawn up and are presented below (in sections 4.4.1 and 4.4.2; with regard to internal validity, these factors are also presented in tabular form; see Table 4.1). (For a detailed discussion of internal and external validity, see chapter 2, section 2.4).

4.4.1 Internal Validity

Table 4.1 overleaf illustrates the threats to

internal validity relevant to the evaluation of the Bangalore project.

Table 4.1

Sources of Internal Invalidity in 4 Schools

	Sources of Invalidity						
School	1	2	3	4	5	6	7
	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality
Bangalore	+	+	+	?	+	?	?
Cuddalore	+	+	+	?	+	?	?
T.Nagar	+	+	+	?	+	?	−
Tiruvottijur	+	+	+	?	+	?	?

− indicates a definite weakness

+ indicates some control

? indicates a possible source of concern

It may be seen from Table 4.1 that there are several potential threats to the internal validity of an experiment. These sources of possible invalidity are taken from Campbell and Stanley's (1963) set of categories. With reference to the comparison of CTP and Structure-based programs (and to Table 4.1), each source of invalidity is now discussed in turn.

History

This refers to events occurring during the course of the experiment which were not controlled for and which could amount to a competing explanation for the results obtained. Examples from language teaching are that students may have practised English at home with their parents (Hatch and Farhady, 1982, p.7) or, at least, outside the classroom (Long, 1984, p.411). To become a plausible rival hypothesis, this should have happened to most of the students in a group (Campbell and Stanley, 1963, p.177). Specific events occurring between the inception of CTP teaching and the evaluation tests are not thought to have influenced the experimental and control groups differently. Although certainty is impossible, there is no available information which indicates that history could constitute a rival hypothesis to the effect of the experimental program.

Maturation

This refers to developmental changes operating

within students independently of the experimental variable, like simply growing older. An example relating to language learning provided by Long (1984, p.412) is that ESL learners living in an English-speaking environment may develop attitudes as a result of that experience, which may in turn lead to increased motivation and higher achievement. With regard to the current study, any processes that may have had a bearing on the test results, such as cognitive development and increase in age, were equally relevant to both experimental and control groups. Maturation is not, therefore, considered a threat.

Testing

This refers to the effect of taking a pretest upon the scores of a posttest. Since no pretest was administered, this factor could not have had a bearing on results of the CTP study.

Instrumentation

This refers to changes in the calibration of the measuring instruments or changes in the scorers used which may influence the results obtained. This may be the case with subjective tests (e.g. essay), which may be marked more leniently by one marker than another, or by the same marker differently at different times. In the case of the current study, the tests used were amenable to more or less objective assessment, and only one scorer

was used throughout. It is not apparent that this factor could have posed a serious threat.

Statistical Regression

This refers to the potential for misinterpretation of results which occurs if comparison groups have been selected on the basis of their extreme performance, since regression towards the mean may tend to subdue the achievement of especially able students and enhance the apparent achievement of especially slow pupils. This is ignored by Long (1984) and by Hatch and Farhady (1982), and may be disregarded for the CTP inquiry since the groups were not selected on the basis of extremes in performance.

Selection

This refers to the differential selection of respondents for the comparison groups which means that the groups may not be equal when the experimental treatment begins. Obviously, this would confound obtained results.

With regard to the current CTP inquiry, pupils were not randomly assigned to experimental and control programs. Campbell and Stanley (1963, p.185) insist that randomisation is the 'only' and the 'essential' way of ensuring the initial equivalence of groups. Not only were the comparison groups not randomly constituted, but also the control groups were not even designated as such until the evaluation was set up.

In spite of this inauspicious beginning, it was worth adopting the peer groups in each school as controls because pupils were not assigned to classes on any differential basis. Instead they were assigned on a basis of first come-first served. Moreover, the CTP team did not try to select a class that appeared to them to be better, but simply took whichever class was made available to them.

It was possible that parents who were more interested in their children's education would apply first, and that these children would come together in the first class to be formed. However, it was established that according to the headmasters and headmistresses involved at the 4 schools, there was no difference between the classes in any year.

The only exception here is the Bangalore school where there were only two classes in each year, and the school was divided into two sites, one class being taught at one site, the other at the other site. The subjects who formed the CTP class were thought to be drawn from a noticeably poorer environment. Therefore, in the Bangalore school, selection bias could be a plausible rival hypothesis, and the experimental group may have been disadvantaged. An attempt was therefore made to establish equivalence by obtaining results of subjects' end-of-year performance in all disciplines (except English) prior to their second year of English. Out of a

total possible score of 400, 46 control pupils achieved a mean of 217.5, and 42 experimental pupils, 196.29. Although the direction of the difference clearly favours the control group, the difference was not statistically significant at the .05 level of significance.

In Cuddalore, the same information was sought but was only partially available. Data was available for only 16 out of 33 control pupils; their mean was 159.31. 34 experimental pupils attained a mean score of 154.18. The difference is slight and not significant at the .05 level of significance. In Tiruvottiyur, no record existed, and in T. Nagar, the reshuffles (referred to below under 'mortality') made the collection of such data largely meaningless.

Another procedure often used when groups are non-random is to administer pretests and make adjustments for differences at the analysis stage through, for example, analysis of covariance. This is an extremely controversial option (see Cronbach, 1982, for a detailed discussion of the dangers) but as has already been noted, the evaluation was not considered at the design stage so no pretests were given.

To sum up, the comparison groups were not formed on a differential basis, and from the limited evidence available, the groups appear to have been equivalent. What took place in all four schools was a form of natural sampling. Although this cannot substitute for random sampling based on probability rules and using, for

example, random numbers, the haphazard natural sampling that occurred in the schools in question seems likely to have resulted in a chance distribution. So while doubt remains, especially in the case of the Bangalore school, it is not sufficient to require total mistrust of obtained results. (Hence the question mark seems appropriate for this factor in Table 4.1).

Experimental Mortality

This refers to a differential loss of pupils from the comparison groups which would affect the results obtained. In the CTP study, the drop-out rate was about equal for 3 out of the 4 schools, but the fourth - T. Nagar - is problematic in this respect. In an attempt to maintain the original constitution of the group, the CTP team made a request to the headmistress that the experimental group be preserved as far as possible. (Normally, the pupils who fared particularly badly would be required to repeat a year, which meant that each new academic year, a class would lose several students and gain several others who were repeating the year above. The CTP team took the view that this would have an adverse effect on CTP teaching and sought to minimise the difficulty of being confronted with a small group of pupils who had not been taught by CTP methods within the larger group who had. Thus, over the course of the CTP teaching, the experimental group lost only 3 subjects and gained 5. The control group, by contrast, lost 10 and

gained 13.

The position, then, was that the experimental group retained its poor students, but was not overburdened with poor students from an older class; and vice-versa for the control group. The teachers involved were of the opinion that poor students repeating a year are less burdensome than poor students who are allowed to continue. If this is so, the experimental group may have been unnecessarily disadvantaged, but clearly any comparison between groups becomes fraught with difficulty.

The position is further complicated by differential attrition from the evaluation tests. In T. Nagar, an average of 10 pupils was absent for the tests compared to an average of 2 for the control group. Worse still, in the experimental group, a total of 26 pupils took some of the tests, but only 10 took all of the tests. Thus, 16 CTP pupils missed at least one of the tests, compared to only 7 in the control group.

It is clear that both of these factors render the T. Nagar results virtually uninterpretable and although the results will be reported in the following section (4.5), they will not influence the rejection or acceptance of the null hypotheses. Mortality could easily constitute a plausible rival hypothesis (hence the minus sign in Table 4.1 under 'Mortality').

4.4.2 External Validity

Randomisation removes most of the threats to internal validity, and it is possible to state with moderate certainty when an experiment is internally valid or where relative uncertainty exists. Thus it is feasible to summarise internal validity considerations in tabular form, as in Table 4.1. The same does not hold for external validity. External validity is only ever a question of degree; thus a table (such as Table 4.1) might easily be full of question marks. So rather than attempt to tabulate the factors that may jeopardise external validity, they will simply be discussed in turn.

In the language of analysis of variance, Campbell and Stanley (1963) note that while internal validity relates to main effects, external validity can be construed as interaction effects involving the program and some other variable(s). As Snow (1974, p.270) observes, "interactions limit generalisation of treatment effects". These interactions potentially confine generalisability to some undesirably limited set of circumstances, and do not permit statements about a larger universe or population that all researchers would presumably wish to address.

The factors affecting external validity put forward by Campbell and Stanley (1963) are rather few and are afforded relatively scant treatment. Bracht and Glass (1968) set out to elaborate these factors and to produce

a more comprehensive (though by no means exhaustive) taxonomy. Later, Snow (1974) built on the models of Campbell and Stanley (1963) and Bracht and Glass (1968) by expanding on the concept of representative design devised by Brunswik (1956). It is mainly on the work of Bracht and Glass (1968) and Snow (1974) that the following discussion is based, though Cronbach's (1982) perceptions (referred to frequently in chapter 2, section 2.41) are also relevant.

Only two main headings will be used: population validity and ecological validity.

4.4.2.1 Population Validity

Population validity is concerned with the degree to which the results of an experiment involving certain subjects are generalisable to other subjects.

Kempthorne (1961) made a distinction between the experimentally accessible population and the target population. It might be wished, for example, to generalise to a target population which comprises all 1st grade EFL learners in French schools; the population accessible to the experimenter might be all of the schools in one city. From the schools actually accessible to the experimenter, a stratified random sample may be chosen for whatever treatment is planned. Results could be quite properly generalised from the sample to the accessible population through inferential statistics;

what holds for the randomly constituted sample holds for the population it was drawn from. However, the second generalisation - from accessible population to target population cannot be made statistically, as the city's schools may not be representative of all French schools, however well stratified the samples were (for instance to include private schools, state schools of different kinds, Catholic schools, etc.). Because the accessible population was not randomly drawn from the target population, no statistical generalisation is possible.

This is problematic for educational research because no researcher is ever able to draw freely from such a large and varied population, and so the second generalisation is effectively denied. It is only a few projects that are massively funded which can even hope to sample from an experimentally accessible population of any statistically viable size. What nearly always happens is that the experimenter takes what few schools are made available (through personal contacts and goodwill) and carries out the planned experiment in them.

The immediate effect of this is that the sample becomes the experimentally accessible population, and no statistical generalisation can be made. Results are limited to the schools which took part in the experiment.

In the Bangalore project, the CTP team managed to obtain the cooperation of 8 schools at different times, and four which were available at the time of the current evaluation. The four schools - the sample - form the

experimentally accessible population, and statistically, at least, the results obtained are confined to that small population.

The only option open to educational researchers in general and to the current inquiry in particular is informed guesswork, that is, to generalise to subjects like those observed (a procedure recommended by Lindquist, 1953; Cornfield and Tukey, 1956; Bracht and Glass, 1968; Snow, 1974; and Cronbach, 1982). In order to accomplish this, it is essential to have as full and relevant a description as possible of the experimentally accessible population and of the apparent threats to external validity.

The students in the 4 CTP schools were all aged between 9 and 12 years. Therefore, a conservative generalisation would only be possible to children of the same age. It is possible that older pupils would have responded differently; e.g. they may have demanded practice with grammatical models and reacted adversely to attention being focused on meaning.

There were classes consisting only of boys (Tiruvottiyur), only of girls (Cuddalore), and mixed (Bangalore and T. Nagar). If the CTP pupils had only been of one sex, strictly speaking, results could not readily be generalised to the other sex. In the event, sex would appear not to be a limiting factor.

It is not known how the CTP pupils rated on I.Q.

tests as this information was not accessible to the author. There is no measure of how the pupils stood in relation to their peers within the same states (Tamil Nadu and Karnataka) or within the country on any academic measure. From the perspective of population alone, generalisation even to other pupils in the same states is groundless (though, as will be seen below, ecological information provides some basis for such generalisation).

It would only be moderately helpful, even if all of the information were available, simply to note the age, sex and I.Q. of subjects. It would be more useful to provide measures of such "personological" (Bracht and Glass, 1968) variables as intuition or research deem relevant.

One variable that has received considerable attention in the literature has been the possibility of interactions between educational programs and the concept of field-dependence and -independence developed by Witkin and his many collaborators (see Witkin et al 1977). An earlier proposal by the author to investigate the effect of cognitive style on learning via the CTP and structure-based models proved not to be feasible and had to be abandoned. It was not possible, therefore, to gather information on this variable.

With regard to population characteristics, then, comparatively little is known. Therefore, extrapolation to other populations would be especially precarious in terms other than age and sex.

4.4.2.2 Ecological validity

Ecological validity refers to the extent to which the results of an experiment involving certain situations are generalisable to other situations. (Situations include treatments, duration, institutional environments, teachers, socioeconomic levels, and so on).

In the same way that statistical generalisation is usually denied to educational research with regard to subjects, so it is normally denied to situations. Where, for example, accessible settings are not randomly drawn from the target universe of settings, we are dependent on accurate and complete descriptions which permit extrapolation from some sites to others. As with population, with ecology it is also possible to take steps to maximise the generalisibility of experimental results. The features that make the results of the Bangalore project more or less generalisable from the perspective of ecological validity are now considered.

(i) The teachers

One factor of the CTP which inhibits generalisation is that most of the teachers in the experimental groups were not regular teachers in the schools. Campbell and Stanley, more than 20 years ago, identified this as a potentially limiting factor in research on teaching:

> the present authors are gradually coming to the view that experimentation within schools must be conducted by regular staff of the schools concerned, whenever possible, especially when

> findings are to be generalised to other classroom situations. (1963, p.191).

In chapter 5, the levels of implementation of the regular teachers was found to be substantially (though not significantly) different from those of the non-regular teachers. If the project had been carried out only with regular teachers, it is conceivable that the results obtained would have been lower. Therefore, it is impossible to place confidence in any assertion that the CTP results are generalisable beyond the confines of the original project, at least as far as the teacher variable is concerned.

(ii) School or 'laboratory'?

A feature of the CTP that should enhance external validity is the fact that the project took place in regular schools, in which there was no artificial control of the environment, no attempt to regulate teaching conditions and student behaviour in some abnormal fashion in order to increase internal validity. It has been argued in Chapter 2 that researchers such as Freedman (1976) manipulated their experiments to an extent where their results were restricted to the 'laboratory' they had created, and to the past tense (i.e. in these particular circumstances, at this particular time, these results were obtained). In the Bangalore project, there were real teachers (though not necessarily regular), real treatments (i.e. not 'canned' or recorded), real extraneous variables, such as high noise levels produced

by traffic from the streets, crowded classrooms, and so on. This permits some confidence that other schools with similar conditions might be able to reproduce the results. The natural setting assists in extrapolation.

(iii) More than one school involved

The fact that there was some attempt at replication in the project also serves to enhance extrapolation to other schools and circumstances. 4 schools took part in the evaluation and these were far apart geographically. Two were in different parts of Madras (one central, one on the outskirts in a fishing community); one school was in the same state (Tamil Nadu) as the Madras schools, but was situated in a small town (Cuddalore); and one school (the Bangalore school) was in a different state (Karnataka). Three of the schools were mission schools, and one, though a state school, was of a higher standard than the corporation schools. It is possible to extrapolate findings from these 4 schools to other similar schools with some small degree of confidence simply because there were 4 schools and not one. Extrapolation to another mission school in Bangalore would not require an act of faith, as the extrapolation would not be travelling a long distance (by this is meant that where similar conditions obtain, generalisation is rendered less tenuous). By contrast, extrapolation to a corporation school in Madras would be extravagant because it is known that the project could not be set up satisfactorily in a corporation school in Bangalore (as

reported in section 4.3.1 above).

(iv) Systematic monitoring

A problem consequent upon the late attention to external evaluation is that the lessons in the project schools were not systematically monitored and only a small number of haphazard audio and video recordings were made (these are analysed in chapter 7). The upshot is that descriptions of the CTP treatment are casual, which in turn means that anyone attempting to replicate the project teaching would be inadequately equipped to do so. One of the perennial problems of program evaluation (see chapters 1 and 2) has been the lack of systematic attention to the independent variable. An attempt to replicate the project would have to depend on the lesson reports put out by the RIE, the descriptions of the project by the director, on the accounts of the teachers (see chapter 5), and on the transcipts of the recordings (chapter 7). This is less substantial and complete than is ideal for generalisation purposes, because the best available descriptions could be interpreted variously.

(v) Novelty and disruption effects

Cronbach (1963) noted that comparative studies of competing curricula were often difficult to interpret because of uncertainty about whether the apparent superiority of one program is real or whether it is due to the innovative effect. The CTP, as an innovative program, may, for example, have produced greater

motivation among teachers and pupils primarily because it was novel. This is more likely in an environment where change is rare. In the project schools, English had been taught via structure-based methods ever since the 'Madras snowball' (Smith 1962), when 30,000 teachers were retrained. The CTP required different teacher - pupil role relationships in that it encouraged students to speak out and teachers to refrain from controlling all of their utterances. In addition, as has already been noted, the CTP teachers were mostly outsiders to the project schools, which would have increased the strangeness of the innovation.

Against this, however, is the fact that the CTP was not a short-term program; it ran for an academic year (in Tiruvottiyur) or longer. It could be argued that lengthening the duration of a program would cause the novelty to wear off. Strictly speaking, certainty is never possible, and even long-term programs may have succeeded because the initial novelty effect had led to more permanent skills in the experimental subjects. However, given that generalisability is a matter of degree, it is reasonable to suggest that the duration of the CTP was sufficient to reduce the likelihood of novelty effects confounding extrapolation.

The opposite of the novelty effects is the possibility of disruption effects, i.e. when an innovation is sufficiently different to what has gone before, teachers and students may have difficulty

adjusting and this may tell against the program. Once again, since the CTP ran for long periods, any disruption effects should have been overcome. (There is always the possibility, as Bracht and Glass (1968) observe, that novelty and disruption effects may cancel each other out).

The Hawthorne effect is also a factor influencing generalisation, since subjects' awareness of participating in an experiment may precipitate behaviour which would not occur in a setting not perceived as experimental. One way around the Hawthorne effect is to make the control program 'experimental' too, but this was not done in the Bangalore project. CTP classes could not have been unaware of being guinea-pigs, as special teachers were assigned to them, the experimental treatment was markedly different from the type of teaching the pupils had ever experienced, and the lessons were heavily observed not only by members of the CTP team, but also by the headmasters/headmistresses and a number of visitors from Great Britain.

It is difficult to gauge just how serious a threat the Hawthorne effect is. Gage (1978) cites the only two reviews of the appropriate literature he can find, both of which conclude that the Hawthorne effect probably does not contaminate results to the extent that some researchers have claimed. The placebo effect in medical studies, whereby subjects' enormous faith in the power of

treatment actually produces improvement, is known to occur very frequently (Rosenthal and Frank 1956); similarly, in psychology it has been documented that subjects who are aware of taking part in an experiment tend to react extremely compliantly and diligently, such is their regard for scientific investigation (Orne 1962); in education, and in language teaching in particular, it remains a matter of speculation.

If the Hawthorne effect is a serious threat, then even the attenuating influence of time may not have dissipated the effects of periodic visits by distinguished personnel. Extrapolation to a non-experimental situation would therefore be tenuous.

(vi) Duration of treatment

Snow has this to say about duration:

> Most school learning situations are extensive in time. No doubt there are some purposes for which short experiments are sufficient or even desirable. But most generalisations about school learning need to be built on research using substantial samples of learning time. (1974, p.281).

As has been argued in chapters 1 and 2, many of the studies carried out in education depend on very small samples of learning time. Because of this, the findings of such studies cannot be extended with any confidence beyond the time scales used in the studies. This means they have very little to say about school practice.

The CTP was a long-term study and, as such, lends itself more readily to extrapolation to circumstances where learning takes place over time (i.e. most secondary

schools, but not, for example, 4-week summer courses).

(vii) Multiple-treatment interference

This comes into effect whenever multiple treatments are applied to the same subjects, because the effects of prior treatments are not usually erasable or assessable.

In 3 of the 4 project schools, only one treatment was applied to each group, that is to say, the experimental groups were exposed only to CTP teaching from the zero level, and the control groups only to structural teaching. Therefore, there seems no possibility of multiple-treatment interference.

The exception is the Tiruvottiyur school, where subjects in both groups had been exposed to 3 years of structural teaching before the experiment began. In the fourth year, the experimental group was taught by CTP methods while the control group continued with the structure-based program. Thus, it is difficult to judge how far the CTP group's results were due to the prior treatment or to the experimental treatment. Clearly, then, generalisation from the Tiruvottiyur school would be inhibited by concerns about which program brought about the obtained results.

(viii) Treatment X Time Interactions

One all-pervasive threat to the external validity of an inquiry is not mentioned by Campbell and Stanley (1963), Bracht and Glass (1968), nor by Snow (1974). That is the fact that generalisations about man and his

institutions are subject to fluctuations over time. It was left to Cronbach (1975), to enunciate clearly just how vital it is that behavioural scientists appreciate this. "Generalisations decay" (Cronbach 1975, p.122), he wrote, adding that

> the trouble, as I see it, is that we cannot store up generalisations and constructs for ultimate assembly into a network. (1975, p.123).

Cronbach notes that while those physical scientists examining such phenomena as atoms and electrons can be sure that their propositions have relatively long half-lives, propositions relating to man and his creations are constantly influenced by changes in local conditions, and have, therefore, relatively short half-lives.

The implication for the CTP is that the results may only be generalisable for a short period. Just how long, it is impossible to say, but it will be influenced by such factors as changes in language policy, changes in the local area brought about by new industries or the introduction of new social customs, just to mention a few of the limitless possibilities. In other words, an endless range of possible interactions could arise merely as a function of time passing. (It is for this reason that evaluations are unusable unless timely; see Cronbach 1982).

The CTP was helped by being set in schools rather than 'laboratories', by operating in more than one site, and by being of relatively long duration. This does not of itself mean that all possible interactions can be

predicted. But if we take Cronbach's view that the only prudent aspirations for any social science amount to pinning down "the contemporary facts" (1975, p.126), then at least a project which is largely a naturalistic study permits an investigator to look behind the test scores and attend to current interactions, i.e. to interpret in context (this is the purpose of chapters 5, 6 and 7). A study set up as a true experiment, by contrast, systematically cleansing the environment of all the influences which will obtain in any real world setting (e.g. Freedman 1976) rules out interactions as unwanted extraneous variables. Those very interactions, however, could easily wipe out the main effects (Cronbach and Snow 1981).

4.5 Results

The results of the tests in the four schools that took part in the evaluation are presented as raw scores in Appendix 3. The summary statistics are displayed in tables 4.2 to 4.6 below.

The summary statistics include the number of items in each test (after item analysis) (No. of items), the mean ($\bar{x}$) and standard deviation (SD) for each group, the number of subjects (N), Kuder-Richardson 21 reliability coefficients (r), t-values (t), degrees of freedom (df), and two-tailed significance (p). In the column headed 'p', 'E' indicates that the experimental group did

significantly better, 'ns' that there was no significant difference, and 'C' that the control group did significantly better; * indicates the .05 level of significance, ** indicates .01, and *** indicates .001. Under the column heading 'No. of items', the number for the same test type varies between schools. This means that either different tests were used at the different schools, or that the same tests were used and trimmed differently after item analysis.

Table 4.2

Test (i): Structure

School	No. of Items	$\bar{x}$	SD	N	r	t	df	p
BANGALORE:								
Control	14	10.27	2.22	48	.60	4.50	88	C***
Experimental		8.07	2.36	42				
CUDDALORE:								
Control	14	8.31	3.28	29	.80	4.82	61	C***
Experimental		4.53	2.85	34				
T. NAGAR:								
Control	19	8.63	3.64	30	.72	3.49	48	C***
Experimental		5.15	2.97	20				
TIRUVOTTIYUR:								
Control	15	10.07	3.03	60	.75	2.73	114	C**
Experimental		8.41	3.44	56				

Table 4.3

Test (ii): Contextualised Grammar

School	No. of Items	$\bar{x}$	SD	N	r	t	df	p
BANGALORE:								
Control	16	8.17	3.95	47	.78	1.97	82	ns
Experimental		9.76	3.17	37				
CUDDALORE:								
Control	15	3.38	3.11	29	.90	2.38	61	E*
Experimental		6.03	5.15	34				
T. NAGAR:								
Control	12	5.64	2.95	33	.75	1.65	53	ns
Experimental		4.32	2.72	22				
TIRUVOTTIYUR:								
Control	13	6.11	2.45	53	.65	1.08	104	ns
Experimental		5.55	2.85	53				

Table 4.4

Test (iii): Dictation

School	No. of Items	$\bar{x}$	SD	N	r	t	df	p
BANGALORE:								
Control	24	15.11	6.17	46	.88	3.01	83	E**
Experimental		18.67	4.21	39				
CUDDALORE:								
Control	28	8.79	8.74	29	.95	1.19	61	ns
Experimental		11.38	8.25	34				
T. NAGAR:								
Control	21	14.60	6.70	30	.93	1.55	48	ns
Experimental		11.70	5.82	20				
TIRUVOTTIYUR:								
Control	28	18.34	7.05	56	.91	0.70	105	ns
Experimental		19.29	6.89	51				

Table 4.5

Test (iv): Listening/Reading Comprehension

School	No. of Items	$\bar{x}$	SD	N	r	t	df	p
BANGALORE:								
Control	28	9.20	6.36	46	.92	6.65	82	E***
Experimental		18.26	5.98	38				
CUDDALORE:								
Control	27	3.52	5.14	29	.97	9.01	61	E***
Experimental		18.03	7.88	34				
T. NAGAR:								
Control	24	10.03	7.19	30	.92	1.29	48	ns
Experimental		7.65	5.93	20				
TIRUVOTTIYUR:								
Control	26	11.26	5.87	60	.91	2.78	114	E**
Experimental		14.73	7.38	56				

Table 4.6

Test (v): CTP Task-Based

School	No. of Items	$\bar{x}$	SD	N	r	t	df	p
BANGALORE:								
Control	25	12.02	5.51	48	.90	6.42	88	E***
Experimental		19.26	5.00	42				
CUDDALORE:								
Control	16	3.39	3.47	33	.89	3.00	64	E**
Experimental		6.76	4.89	33				
T. NAGAR:								
Control	24	9.31	6.36	32	.90	2.72	50	E**
Experimental		14.00	5.16	20				
TIRUVOTTIYUR:								
Control	21	11.21	4.99	58	.89	2.00	111	E*
Experimental		13.74	5.83	55				

Tables 4.2 to 4.6 may be summarised as in Table 4.7 below, focusing solely on patterns of significance:

Table 4.7

Patterns of Significance for 4 Schools and 5 Tests

School	Test 1	Test 2	Test 3	Test 4	Test 5
Bangalore	C***	ns	E**	E***	E***
Cuddalore	C***	E*	ns	E***	E**
T. Nagar	C***	ns	ns	ns	E**
Tiruvottiyur	C**	ns	ns	E**	E*

'*' indicates the .05 level of significance

'**' indicates the .01 level of significance

'***' indicates the .001 level of significance

'ns' indicates no significant difference

The concern with test content bias was slightly lessened by the results. Although in 5 out of 12 possible results on tests (ii), (iii) and (iv), the experimental group was significantly better, in the other seven, there was no significant difference. This suggests that both groups were able to compete on these measures. This is, however, only a crude interpretation, and needs to be backed up by more rigorous examination (see section 4.5.1 below).

As mentioned in section 4.4.1 above (under 'experimental mortality'), the T.Nagar results would be reported, but would not be subjected to further analysis and they would not influence the acceptance or rejection of the null hypotheses (because of the serious threat to internal validity which makes them virtually uninterpretable).

In the other 3 schools, it can be seen from Tables 4.2 to 4.7 that the control groups performed significantly better on test 1 (the achievement test designed to reflect the teaching that had taken place in the control classes) and that the experimental groups performed significantly better on test 5 (reflecting CTP teaching). This outcome would lead us to reject the first null hypothesis, and argue instead that there is, in fact a difference in the abilities arising from the control and experimental teaching. (Of course, an outcome in which one group performed significantly better on <u>both</u>

tests would also have permitted the rejection of the no-difference hypothesis, but would have had the added benefit of facilitating the comparison and obviating the need for further tests).

For the second null hypothesis to be rejected, it would have been necessary for the experimental groups to perform significantly better on the syllabus-neutral proficiency tests of contextualised grammar and dictation (tests 2 and 3). In the event, the experimental group fared significantly better in one school out of three (Bangalore) on the dictation and in one school out of three on the contextualised grammar test (Cuddalore). All of the other results were non-significant, but the direction of the difference was in the experimental groups' favour in 3 out of 4 cases. Thus, although the null hypothesis cannot be rejected outright, it can be reasonably argued that it can be partially rejected. Thus, instead of 'Acquisition of non-syllabus-based structure is best achieved without focus on form', the formulation 'Acquisition of non-syllabus-based structure can be achieved without focus on form' seems appropriate.

As for the third null hypothesis, its rejection required the experimental groups to score significantly higher than the control classes on the proficiency test of listening/reading comprehension which, as Table 4.5 records, turned out to be the case. The results allow the claim that 'structure acquired without focus on form is more readily available for deployment than structure

learned with focus on form'.

4.5.1 Validity and Reliability of the Tests

While the results described in the above section indicate that null hypotheses 1 and 3 may be rejected, and that null hypothesis 2 may be partially rejected, it is necessary to have some measure of how much trust we may place in the results, and to what extent the tests measured what they claimeed to be measuring. Questions of this sort refer to the twin concepts of reliability and validity.

4.5.1.1 Reliability

The form of reliability that was examined was a measure of internal reliability, using the Kuder-Richardson 21 formula. First of all, item analysis was carried out and the items which failed to discriminate above .30 or whose facility value was outside 25 and 85 were dropped. This meant that tests that were originally of the same length at different schools were trimmed variously. The Kuder-Richardson 21 formula was then applied and the results for each test at each school are reported in Tables 4.2 to 4.6 above. It can be seen that for tests 3, 4 and 5, the coefficient ranges from 0.88 to 0.97; for test 2, it is somewhat lower (0.65, 0.78 and 0.90); and for test 1, slightly lower still (0.60, 0.75 and 0.80). However, in only two cases is it low enough to

account for less than 50% of the variance (0.60 and 0.65).

The reliability coefficients are mostly very high and never disturbingly low (except perhaps in two instances). It may be concluded, therefore, that the results of the tests were consistent, a necessary condition for their interpretation.

4.5.1.2 Validity

Validity is typically divided into 5 types (Davies 1977): (i) face, (ii) concurrent, (iii) predictive, (iv) content and (v) construct. The tests used in the evaluation will now be considered from each of these perspectives.

(i) Face.

This refers to the degree to which a test 'looks right', i.e. the extent to which teachers, pupils and administrative personnel consider that the test has valid properties. This is not susceptible to measurement (though see Nevo [1985] for a recent attempt, and Secolsky [1987] for a response appealing for caution). Face validity is more important for a test which is public and which is to be administered year after year. For one-off tests intended to serve for very specific circumstances, as in the CTP evaluation, it is a less pressing concern. Nevertheless, there is no reason to assume that the tests lacked this form of validity.

(ii) Concurrent

This refers to a form of validity in which the test in question is correlated with another test which has already been validated and is regarded as an acceptable criterion.

No criterion tests were available which could be considered either valid or appropriate to the testing requirements of the present inquiry (at least, none were known to the author). No public tests of English language had been devised in Tamil Nadu or Karnataka for the age and grade levels of the comparison pupils, and even those prepared for higher school grades had not been validated.

(iii) Predictive

Like concurrent validity, predictive validity is established through correlation with another test. This time the criterion test is a test which will be taken at some stage in the future, and the test we wish to validate predicts performance on that future criterion.

It was beyond the scope of the present inquiry to keep a check on students' future performance, even if a suitable criterion could be found, and even if many of the pupils did not disperse. Logistically, no provision was made for a predictive validity study.

(iv) Content

Content validity is usually seen as the most important form of validity for a program evaluation, simply because it is usually desirable to measure the

learning that has taken place under a particular treatment. It requires an analysis of the syllabus or any other document which specifies what material was covered, and a sampling from that specification. As Davies puts it, "the validity is established by an expert appraisal of the test content as a sample of the subject to be learned" (1977, p.61-62).

Relevant to the present evaluation, the intention was to seek content validity for the tests 1 and 5 which were referenced to the structure-based program and the CTP program. Test 1 was intended to reflect the structural syllabus that the control classes adhered to, and to this end, the syllabuses issued by the states of Tamil Nadu and Karnataka were examined and samples of structures were taken from them. Although a structure test had been written before arrival in India, it had to be revised to accord with the syllabuses (and perceived pupil levels), though the choice of format (multiple-choice) was a very familiar one in the control classrooms and was retained.

The input for the CTP test was the documentation issued by the RIE in the form of lesson reports and a tentative syllabus of tasks (in the RIE Newsletters, Bulletins, and Lesson Reports). A sample of tasks was rewritten in the same format but with different particulars (e.g. distances between towns would be different from those in the tasks on the cyclostyled

handouts given to teachers).

The director of the CTP, Dr. Prabhu, affirmed that, in his view, the two tests fairly reflected the content of the comparison groups. Since there is no statistical means of establishing content validity, the opinions of 'evaluators' and 'clients' are probably the appropriate ones for a program evaluation.

(v) Construct

Construct validity is the most complex of the validities, and its function is to defend a proposition about what a test measures. There is no one way of establishing this form of validity and a wide range of research techniques are considered pertinent (Cronbach 1984, p.152-153). All of the other validities contribute to it. This is so because the end goal of all validation is explanation and understanding of what a test tests.

In the evaluation of the Bangalore project, tests 2 and 3 purport to comprise constructs that are central to tests 1 and 5 ('structural accuracy' and what might be construed as 'fluency'). In order to demonstrate that these tests were not biased toward either program, these constructs should both be present.

In section 4.3.4.4, it was argued that test 3 should be fair for pupils who approach it from a perspective of structural accuracy and also for pupils who would adopt an approach based on an understanding of meaning. Similarly with test 3. There is reason to believe that

second (and central) hypothesis stating that acquisition of non-syllabus-based structure is not best achieved without focus on form was partially rejected. The third hypothesis that structure acquired without focus on form is no more readily available for deployment than structure learned with focus on form was rejected.

However, these results cannot be taken at face value. A detailed consideration of the sources of potential internal and external invalidity led to a more cautious appraisal. First of all, the T. Nagar results were dropped from further analysis because the conduct of the project and the examinations in this school rendered interpretation considerably more hazardous than in the other 3 schools.

Even in the other schools, external validity was constrained because little was known about the characteristics of the sample population. Furthermore, the presence of non-regular teachers, novelty, disruption and Hawthorne effects, multiple-treatment interference (at one school) also hamper generalisation and are only partially balanced by the duration of the project, its 'natural' setting, and the fact of replication in different sites.

Quite apart from design problems, the tests, although reasonably reliable, could not be fully validated. Construct validity for the proficiency tests 2, 3, and 4 was particularly difficult to establish,

the content validity for tests 1 and 5 is reasonably high, so it needs to be shown that the constructs that underlie tests 1 and 5 also underlie tests 2 and 3.

As for test 4, since it was expressly designed to take account of fluency, it would be expected to share substantial traits with test 5, somewhat fewer with test 2 and 3, and to have only a weak relationship with test 1.

In order to discover how many underlying traits our tests exhibit, a factor analysis was applied. Beforehand, however, correlation matrices were examined for interrelationships and for indications about whether the 2 anticipated factors seemed likely to emerge. The correlation matrices are presented in Appendix 4, but the same information is presented below in Table 4.8, arranged in such a way as to make the appropriate interrelationships more readily discernible.

Table 4.8

Correlations between 5 tests in 3 schools

Tests	BA	CU	TV	Average
1 and 2	.25	.26	.39	.30
1 and 3	.45	.44	.52	.47
1 and 4	.18	-.10*	.50	.29
1 and 5	.09*	.19	.37	.22
2 and 1	.25	.26	.39	.30
2 and 3	.42	.75	.39	.52
2 and 4	.46	.74	.46	.55
2 and 5	.44	.62	.43	.50
3 and 1	.45	.44	.52	.47
3 and 2	.42	.75	.39	.52
3 and 4	.66	.60	.63	.63
3 and 5	.61	.73	.55	.63
4 and 1	.18	-.10*	.50	.29
4 and 2	.46	.74	.46	.55
4 and 3	.66	.60	.63	.63
4 and 5	.71	.69	.66	.69
5 and 1	.09*	.19	.37	.22
5 and 2	.44	.62	.43	.50
5 and 3	.61	.73	.55	.63
5 and 4	.71	.69	.66	.69

For all correlations, $p < .05$, except those marked *.

BA = Bangalore; CU = Cuddalore; TV = Tiruvottiyur.

It may be seen from Table 4.8 that test 5 correlates highly with test 4, somewhat less so with tests 2 and 3, and very slightly with test 1. This conforms with our expectations.

Test 1 correlates moderately with tests 2 and 3, but lowly with tests 4 and 5, again as expected. The only discrepancy here is that in the Tiruvottiyur school there is a somewhat stronger relationship between test 1 and tests 4 and 5 than our theory would predict. However, this can be explained by the fact that in this school, the experimental group had been exposed to structural teaching for 3 years prior to their CTP year. This would sensitise them more to accuracy than the Bangalore and Cuddalore pupils, and tend to subdue the differences between the CTP and control groups on test 1 that were apparent in the other schools.

Test 4, as expected, exhibits a stronger relationship with test 5 than it does with tests 3 and 2; also, it is only weakly correlated with test 1.

Tests 2 and 3 show a moderate correlation with all other tests, but the relationships are stronger with tests 4 and 5 than with 1, suggesting a possible slight bias in favour of the more fluency-oriented training of the CTP groups.

What emerges from a consideration of the correlation matrices is that the postulated 2 factors may have some substance.

Turning to the factor analysis (FA)(run on version 2 of the Statgraphics PC program), Table 4.9 reports the variables (the 5 tests), the communalities, the factors, the eigenvalues, the percent of variance accounted for by each variable, and the cumulative percentages.

Table 4.9

Communality, Eigenvalue and Variance

Variable	Communality	Factor	Eigenvalue	%Variance	Cum%
Bangalore:					
Test 1	0.28065	1	2.79601	55.9	55.9
Test 2	0.27026	2	1.00104	20.0	75.9
Test 3	0.59642	3	.64727	12.9	88.9
Test 4	0.60333	4	.28373	5.7	94.6
Test 5	0.57334	5	.27195	5.4	100.0
Cuddalore:					
Test 1	0.42761	1	3.14250	62.9	62.9
Test 2	0.70845	2	1.12210	22.4	85.3
Test 3	0.74186	3	.38576	7.7	93.0
Test 4	0.72721	4	.19914	4.0	97.0
Test 5	0.63648	5	.15050	3.0	100.0
Tiruvottiyur:					
Test 1	0.34131	1	2.97511	59.5	59.5
Test 2	0.26807	2	.66052	13.2	72.7
Test 3	0.48538	3	.65402	13.1	85.8
Test 4	0.57103	4	.39450	7.9	93.7
Test 5	0.47951	5	.31585	6.3	100.0

Statistical Graphics Corporation (1986, ch.21, p.22) recommends that "by choosing the highest eigenvalues, you can decide which factors to extract for further analysis". Looking at Table 4.9, it is clear that in the first two schools, much of the variance is accounted for by the first two factors. In the third school, the picture is less clear. Nevertheless, the higher eigenvalues coupled with Hatch and Farhady's injunction that "one should have logical reasons for asking for the number of factors specified" (1982, p.262) still suggest that it would be defensible to extract 2 factors.

Extracting 2 factors, and then putting the obtained factor matrices through a Varimax Rotation, the following rotated factor matrices resulted (presented in Table 4.10).

Table 4.10

Varimax Rotated Factor Matrices for 3 Schools

School	Variable	Factor 1	Factor 2
Bangalore	Test 1	0.07877	0.97138
	Test 2	0.62861	0.26294
	Test 3	0.73435	0.47712
	Test 4	0.89173	0.07963
	Test 5	0.90121	-0.04857
Cuddalore	Test 1	0.07239	0.97666
	Test 2	0.86818	0.20106
	Test 3	0.81704	0.44795
	Test 4	0.91673	-0.22729
	Test 5	0.85638	0.13304
Tiruvottiyur	Test 1	0.67080	0.24913
	Test 2	0.25775	0.96048
	Test 3	0.85403	0.11957
	Test 4	0.82606	0.26817
	Test 5	0.75134	0.26864

An examination of Table 4.10 reveals that as far as the Bangalore and Cuddalore schools are concerned, there are substantial indications that the postulated factors exist. Factor 1, which has been loosely labeled 'fluency', correlates highly with tests 4 and 5 and negligibly with test 1. It loads slightly more highly than anticipated on tests 2 and 3. Factor 2 (labeled 'accuracy') loads highly on test 1, negligibly on test 5 and has a negative relationship with test 4. It exhibits the expected correlation with test 3, though it is slightly subdued in test 2.

The picture that emerges here is that 'fluency' and 'accuracy' account for a large amount of the variance in the 5 tests, and that the loadings are consistent with our anticipations. Also, it would appear that tests 2 and 3 favoured the CTP groups rather more than would have been desirable.

However, when we turn to the Tiruvottiyur school, it is difficult to see any pattern atall. It was expected that the distinctions between the constructs would be less clear than in the other 2 schools (which had both been learning English for 2 years and all of their exposure had been to the CTP). The same tests might be tapping rather different competencies with groups who had been learning English for 4 years (all of it structure-based for one class, 3 years of it structure-based for the other). In view of this, and also because the

eigenvalues indicated that a third factor was as important as the second, the factor analysis was run again for Tiruvottiyur, this time extracting 3 factors. The results are presented in Table 4.11.

Table 4.11

Varimax Rotated Factor Matrix for Tiruvottiyur

Variable	Factor 1	Factor 2	Factor 3
Test 1	0.20223	0.92183	0.19969
Test 2	0.24586	0.18810	0.94505
Test 3	0.69059	0.52754	0.07237
Test 4	0.79010	0.32951	0.22680
Test 5	0.88294	0.05848	0.23684

Table 4.11 shows that when 3 factors are extracted, the Tiruvottiyur scores reveal loadings on factors 1 and 2 largely consistent with the other two schools. However, the third factor is loaded almost entirely on test 2, which is not interpretable in terms of the expectations described above. No clear label suggests itself for this third factor. The lack of an a priori theory comprising a third factor means that there is a danger of opportunistic post hoc rationalisation, so the messiness of the predicament is merely acknowledged.

The confidence that may be placed in this factor analytic study is constrained in any case because the data are not normally distributed. As Woods, Fletcher and

Hughes (1986, p.139) point out, "if the data do not meet the assumption of normality, the results of factor analysis ... should be treated with extra caution". Only one third of the 15 distributions (5 tests by 3 schools) is normally distributed, and the other two thirds deviate significantly from a normal distribution (as calculated using the chi-square goodness of fit between observed scores and expected scores, i.e. between a fitted normal distribution line and the actual data. The results are displayed in Appendix 5).

Factor analysis is a particularly controversial technique, so that even if the derived factors were all perfectly clear and the data were normally distributed, we would still have cause to view the interpretations with caution. Woods, Fletcher and Hughes (1986) and Hatch and Farhady (1982) note the value of factor analysis, but stress the potential for misuse.

First of all, it is possible to select the number of factors in a post hoc manner. Secondly, the solution (unlike principal component analysis) does not give a unique solution; a different form of rotation with a different number of iterations would vary outcomes. Thirdly, the process of attaching labels to factors is fraught with potential error; for instance, what has been labeled 'fluency' above may is only a best guess based on a view of the tests and the teaching programs; it may, in fact, be quite misleading.

Although not all statisticians take the same view,

it is as well to be aware that some tend to dismiss FA out of hand. Chatfield and Collins (1980, p.55) comment that "FA appears to be used very little by statisticians, though it is widely used (and misused) in the social sciences". They devote only a few pages to FA in a book on multivariate analysis, explaining that this is because

> we find ourselves in sympathy with the growing group of statisticians who doubt if FA is worth using except in a few particular types of application. For example, Hill (1977) has said that FA is not 'worth the time necessary to understand it and carry it out'. He goes on to say that he regards FA as an 'elaborate way of doing something which can only ever be crude, namely picking out clusters of inter-related variables, and then finding some sort of average of the variables in a cluster in spite of the fact that the variables may be measured on different scales'. (1980, p.88).

Cronbach (1984), while he affirms that "[FA] and its close relatives are indispensable in reducing statistical data" (p.283), states that it "is no longer so dominant in research on tests; many additional styles of inquiry have ripened" (p.283). Although its influence may have faded in psychology (where it began), it has flourished in the social and educational sciences. Blackith and Reyment speculate that the technique has "persisted precisely because it allows the experimenter to impose his preconceived ideas on the raw data" (1971, p.201). Finally, Davies (1984, p.112) deplores the "present irresponsible use of computer statistical programmes for language test data".

So, returning to the question of construct validity,

has the factor analysis helped reduce the data, or has it merely set up a smokescreen? Obviously, since our data could not fully meet the assumptions of an already controversial technique, increased caution is desirable. All the same, the underlying factors may have the fairly clear interpretations (in 2 schools at least) that have been proposed because the relationships are clear from a visual inspection of the correlation matrices.

The achievement tests rest primarily on content validity. The proficiency tests, dependent on construct validity, appear to have been approachable by both experimental students and control students, though the former may have been favoured. In the final analysis, however, convictions about the construct validity of these tests must remain tentative.

4.6 Summary and Discussion

This chapter has reported the product evaluation which was the first phase of the total evaluation and which was carried out in 1984. The major constraint was that the point of entry was too late for baseline data to be collected or for adequate, systematic monitoring of the CTP teaching to be carried out. In spite of these constraints, 5 tests were devised which were to be used to accept or reject 3 null hypotheses.

The first hypothesis, which was that there is no difference in the language abilities arising from the control and experimental teaching was rejected. The

second (and central) hypothesis stating that acquisition of non-syllabus-based structure is not best achieved without focus on form was partially rejected. The third hypothesis that structure acquired without focus on form is no more readily available for deployment than structure learned with focus on form was rejected.

However, these results cannot be taken at face value. A detailed consideration of the sources of potential internal and external invalidity led to a more cautious appraisal. First of all, the T. Nagar results were dropped from further analysis because the conduct of the project and the examinations in this school rendered interpretation considerably more hazardous than in the other 3 schools.

Even in the other schools, external validity was constrained because little was known about the characteristics of the sample population. Furthermore, the presence of non-regular teachers, novelty, disruption and Hawthorne effects, multiple-treatment interference (at one school) also hamper generalisation and are only partially balanced by the duration of the project, its 'natural' setting, and the fact of replication in different sites.

Quite apart from design problems, the tests, although reasonably reliable, could not be fully validated. Construct validity for the proficiency tests 2, 3, and 4 was particularly difficult to establish,

CHAPTER 5

LEVELS OF IMPLEMENTATION OF THE CTP

5. Levels of Implementation of the CTP

5.1 Rationale for investigating implementation

In Chapters 1 and 2, it was argued that one of the major failings of much evaluation work has been its neglect of the independent variable, i.e., what actually happens to a program when it reaches the classroom.

It has been widely noted in general educational research (cf. chapter 2, section 2.3.1) that innovative teaching programs are often barely distinguishable from so-called traditional programs (Charters and Jones, 1973; Churchman, 1979; Wang, Nojan, Strom and Walberg, 1984). Different methods may have different theories underpinning them, but in practice, they have a tendency to overlap. Sometimes it is evident that there may be more within-program than between-program difference (e.g. Stebbins, St. Pierre, Proper, Anderson and Cerva, 1977). In view of this, we might ask, like Fullan and Pomfret (1977), "when does variation in use become so wide that the original idea is unrecognisable?" (p.358).

The problem has been precisely the same in language teaching, as Chapter 1 makes clear. The independent variable has hardly ever been measured in a satisfactory way in our studies of method, a point which has frequently been made (Long, 1980; Stern, 1983; Richards, 1984). The upshot is that one teacher's Natural Approach could be another teacher's Cognitive Code.

With regard to the CTP, much has been written about precept, along with broad descriptions of practice (e.g. Brumfit, 1984; Johnson, 1982; and especially, Prabhu, 1987; cf. chapter 3). The extensive lesson reports put out by the project team record the content of the tasks and how they were structured, but there is little detail about what happened in the classroom. The intention of the investigation reported in the present chapter is to increase understanding of how the CTP was actually implemented by project teachers.

There are a number of procedures that have been developed to measure implementation (see chapter 2, sections 2.3.2 and 2.3.2.1), primarily involving classroom observation and interviews. For example, Hall and Loucks (1977) used 20-minute taped interviews with teachers, whose responses were rated in terms of 8 Levels of Use according to certain prespecified criteria. The closer teachers came to these criteria, the higher their Level of Use. Stallings (1975) and Wang (1980) chose to concentrate on critical program components, derive scores for each of the variables and then total them across variables. The information was gathered through observation, though Wang supplemented her study with interviews. Wang, Nojan, Strom and Walberg (1984) divided teachers into high, average and low implementors (an arbitrary scale relating to critical program components).

5.2 Development of an implementation measure

Most of the implementation studies reviewed in Chapter 2 and mentioned in Chapter 5.1 above, have been carried out while programs were actually in operation. The Bangalore Project, however, had come to an end before the present study was conceived. Therefore, it was necessary to devise protocols for retrospection rather than for introspection.

5.2.1 Levels of Use

The concept of Levels of Use (Loucks and Hall, 1977) was taken as a starting point. Live interviews asking teachers to introspect obviously had to be ruled out, but it was conceivable that if teachers would write fairly detailed accounts of their experience on the CTP, it would still be possible to infer degrees of use. Cronbach (1982) has called for such accounts to feed into evaluations, and as he sees program evaluation principally as historical research, the perspective fitted the backward-looking nature of any continuing CTP inquiry..

Before asking teachers to produce historical narratives, however, it was necessary to consider some of the difficulties associated with naturalistic study. When a researcher abandons more objective modes of inquiry, involving experimental and control groups, tests, .05 significance levels, and the like, the option is not necessarily to embrace sloppy research. Welch (1983)

has this to say:

> Too often, the choice to conduct [a naturalistic study] is made by default because the investigator is uncomfortable with measurement and statistics. This is a serious problem because I believe they are more difficult to conduct well than experimental studies. (p.101).

The difficulty from our point of view was that if we had simply asked for accounts and then subjectively assigned Levels of Use, the findings of such a venture would be unfalsifiable. It would be impossible for anyone to follow how we had arrived at our decisions, and thus to criticise them or place confidence in them. In short, it would not have been a disciplined inquiry (cf. chapter 2, section 2.4.2).

Cronbach and Suppes (1969) state that the report of a disciplined inquiry

> has a texture that displays the raw materials entering into the argument and the logical processes by which they were compressed and rearranged to make the conclusion credible. (pp.15-16).

That is, the teachers' accounts, the transformation into levels and the coherence of the operation should be transparently confirmable. Thus, in adapting the Levels of Use concept, was important to ensure that these criteria were observed.

The first stage of adaptation was to reduce the number of levels. Hall and Loucks (1977) had identified 8 Levels of Use, but with a population of only 16 teachers and the need to retrospect, it seemed unlikely that such variegated levels could be sustained in the accounts. It

seemed wiser to simplify; eventually, 3 levels were decided upon, though we lacked an a priori research basis for this.

Therefore, so that what subjective processes were involved in this procedure are open to examination, the Levels of Use chart (from Loucks, Newlove and Hall, 1975) is presented in Appendix 6. The reduced levels used for the present study (which we shall henceforth call Levels of Implementation or LIs) are described in Table 1 below.

5.2.2 Levels of Implementation

Table 5.1

CTP Teacher Implementation Levels

Level 1: ORIENTATION State in which the teacher is not fully aware of the CTP, nor how to use it, nor what its effects might be.
Decision Point A: A routine pattern of use is established
Level 2: ROUTINE State in which the teacher's awareness of the principles and methodology of the CTP is well-developed, and in which his/her use is relatively stable.
Decision Point B: Begins to explore possible modifications of the CTP
Level 3: RENEWAL State in which the teacher is aware of the strengths and weaknesses of the CTP and is consciously seeking modifications that will benefit the learners.

It was judged that fairly tangible distinctions that would be more accessible to retrospection could be based on the notions of (a) struggling to come to terms with the CTP, (b) operating comfortably with the CTP and (c) looking for ways of improving the CTP. The Levels of Implementation given in Table 1 reflect these 3 notions.

In collapsing the levels from 8 to 3, the difficulties of retrospection were not the only considerations. It can be seen from Appendix 6 (the Levels of Use chart) that Levels 0, 1, and 2 relate to non-use and pre-use. In our study, only teachers who had used the CTP were approached, so these levels were irrelevant. Also, there seemed little to be gained from a subtle distinction between mechanical and routine use.

Furthermore, it was judged relevant to separate teachers who were unquestioning in their acceptance of the CTP principles and practice and those who retained their independence. Prabhu (1987, p.103) states that his view of "the project - and of pedagogic innovation generally" is not that teachers carefully carry out a set of allowed procedures, but that they should select and modify as their sense of plausibility dictates. It was therefore pertinent for the Levels of Implementation criteria to distinguish those with a more disciple-like approach from those who were innovative in Prabhu's terms. Hence the Level 2 and Level 3 division.

Level 1 was needed to accommodate the possibility

suspected by visitors to the project (e.g. Brumfit, 1984) that typical teachers tended to revert to a focus on form. That is to say, that they had not fully come to terms with the demands of the CTP.

The distinctions between levels could only be effective if they were conveyed in operational language. The operational description is presented below in section 5.2.3.

5.2.3 CTP Teacher Implementation Categories

The CTP Teacher Implementation Categories, based on a close reading of the CTP literature and the Levels of Use chart (Appendix 6), are as follows:

KNOWLEDGE: What the teacher knows about the nature of the CTP.

Level 1: Has only limited general knowledge of the CTP.

Level 2: Has sufficient knowledge of the CTP for appropriate and stable use.

Level 3: Has sufficient knowledge to evaluate the use of the CTP and to seek modifications.

ACQUIRING INFORMATION: The teacher solicits information about the CTP in a variety of ways, including discussion, review of published descriptions and commentaries, sharing plans and problems.

Level 1: Makes little attempt to find out more about the CTP; discusses the CTP only with the director; discusses only discipline and classroom

management; asks for ready-made materials.

Level 2: Reads the Newsletters and Bulletins relating to the CTP; attends seminars; discusses with other CTP teachers; discusses the effects of the CTP and the development of own materials; observes other CTP classes.

Level 3: Tries to find out ways of improving the CTP; discusses with both CTP and non-CTP teachers; discusses possible modifications of the CTP; compares strengths and weaknesses of of own (and others') teaching and seeks modifications to improve pupil learning.

ASSESSING: Examines the effects of use of the CTP; this could be an informal, mental assessment or actual data collection.

Level 1: Only informal reports of the impact of the CTP on students' attitudes and linguistic performance.

Level 2: Checks the impact of the CTP through in-house achievement tests.

Level 3: Examines the strengths and weaknesses of the CTP through some form of comparative testing (not just achievement testing).

PERFORMING: Describes personal use of the CTP.

Level 1: Does not develop own materials; perceives CTP as requiring a great deal of time and effort; finds the transition from structural to CTP teaching difficult; is confused about the treatment of

error; evinces a tendency to focus on language form; simply goes through ready-made tasks with little sense of planning.

Level 2: Develops own materials with reference to a 'model'; feels no great effort or stress in implementing the CTP; puts into practice perceptions of CTP principles out of a sense of conviction, or discipleship, or concern about experimental contamination; treats error according to published CTP perceptions; planning involves constant modification of the challenge level of the tasks based on daily feedback from students.

Level 3: Develops own materials, consciously deviating from the 'model'; develops pedagogic procedures, introducing own ideas; tries out modifications of the CTP to improve pupil learning.

It can be seen from the Levels of Use chart in Appendix 6 that there were 7 categories elaborating behavioural indices; these are: Knowledge, Acquiring Information, Sharing, Assessing, Planning, Status Reporting and Performing. Before the teachers' accounts started coming in, it was impossible to predict how much information they would include. In view of this, it was difficult to guess whether or not there would be sufficient breadth and depth to warrant 7 categories of

operational indices (especially since these categories overlap). Prior to data collection, therefore, 7 categories were still anticipated, although with the understanding that the data might require a reduction in scope.

In the event, although some accounts were reasonably detailed, it proved virtually impossible to identify in a teacher's statement about classroom practice, for example, whether there was a trace of Status Reporting or perhaps of Planning, or whether only Performing could be inferred. Similarly, it was difficult to sustain the distinction between Acquiring Information and Sharing because, for example, attending a CTP seminar could be viewed as both. In short, the framework of 7 overlapping categories proved to be too cumbersome for our data.

As a result of this awareness, Acquiring Information and Sharing were collapsed (with the former label being retained). Also, Planning, Status Reporting and Performing were combined (with the latter label being retained). This left us with 4 categories: Knowledge, Acquiring Information, Assessing and Performing.

The actual indices of behaviour in the Levels of Use chart are ostensibly generic, but though much might have been learnt from a generic instrument, the potential for loss of relevant information outweighed other considerations, and the indices were written to be CTP-specific, using knowledge gained from familiarity with

the literature on the CTP and discussions with CTP participants and other interested professionals. For example, where the Levels of Use categories at Level 6 refer to 'replacement' of the innovation, we have avoided the term because we are more interested in determining degrees of 'modification' (for the reasons mentioned in section 5.2.2 above, viz., Prabhu's concern that teachers should operate with a sense of plausibility rather than of adherence).

As has already been mentioned, the Levels of Use interviews probe teachers' current use of an innovation. Not only that, but the interviews are carried out on several occasions during the life of a project. In this way, a longitudinal profile is constructed. Given, once again, that retrospection may well be less reliable than introspection, it was clear that CTP teachers could not reasonably be expected to differentiate several chronological stages of their experience; fine temporal distinctions might be blurred. Therefore, it was essential, if the developmental aspect of the inquiry was to be retained, for demands on recall to be less subtle. Teachers were, therefore, asked to distinguish between <u>early</u> and <u>late</u> use (see the Introduction to the teachers' accounts that was sent out to each teacher; section 5.3.1, below). In the event, (as we shall see in section 5.4.1.3, below), only one teacher made clear distinctions between early and late use, so the developmental aspect

had to be abandoned in the analysis.

5.3 Data collection procedures

5.3.1 Cover letter and guidelines for teachers

A cover letter explaining the nature of the inquiry was sent to each teacher who had participated in the project. (This cover letter also asked teachers to complete the Stages of Concern questionnaire of Chapter 6, section 6.3.2.1). With reference to the request for teachers to write accounts of their experience, the letter asked respondents to "provide as detailed an account of 'what it was like' as current demands on your time allow." Confidentiality was assured (in the event of comments being cited in publications), it being stated that all teachers would henceforth be referred to as 'Teacher A, B, C, etc.'

Teachers were asked in the interests of timeliness to return their accounts (and the SoC questionnaires) within a month of receiving them. They were asked to send them via Dr. Prabhu, who had elected to co-ordinate the data collection from Madras (some of the teachers could not easily afford the postage costs).

It was anticipated that teachers would respond in varying degrees of detail. Most of the teachers were known to me personally from a visit in 1984 (to carry out the testing phase of the evaluation, reported in chapter 4), and it seemed likely that some would very

conscientiously write at great length, providing a wealth of information, and that others might be less inclined to go to such trouble. In order to secure a minimum of information from as wide a spectrum of participating teachers as possible, it was considered appropriate to send, along with the cover letter, a set of guidelines, indicating the kind of issues that might be addressed. It was expected that this would have the effect of moving some to respond who might otherwise not have done so, as it would be possible simply to respond to the probes without spending a lot of time wondering what to say. It was also expected that such guidelines would serve to jog memories, and open up half-forgotten areas. Finally, it was feared that guidelines would have a constraining effect, and that teachers would not perhaps mention what was salient for them if it was not included in the probes. The trade-off seemed to be that while more responses might be forthcoming, the nature of those responses might be prejudiced. In order, therefore, to offset the possibility of constraint, an introduction to the guidelines was drafted and sent to all teachers. The introduction encouraged teachers to ignore whatever they wished and to add information that the guidelines had neglected. It should be stressed, then, that it was made clear to teachers that they were free to select their own questions for the accounts and to ignore the guidelines if they so wished.

Both the introduction and the guidelines are now set

forth. (The guidelines were drawn up with constant reference to the Levels of Use categories of Loucks, Newlove and Hall, 1975, and to the available literature on the CTP).

Teachers' accounts: Introduction

Only you and the other teachers who were actually involved in the CTP really know what was at stake on a day-to-day basis in the classroom. If teachers and researchers elsewhere are to benefit from your experience, it is important for there to be a record of it. Therefore, a detailed historical narrative written by each of the CTP teachers would be of great value.

To help shape these narratives, 7 broad categories are suggested (knowledge, acquiring information, sharing, assessing, planning, status reporting, and performance), and a number of issues that might arise within each category are set out in the guidelines below. Obviously, you will find more to write about on some issues than others; also, you may find that there is considerable overlap; so please feel free to ignore areas that seem irrelevant to you personally and to add others.

The important thing is to write in as much detail as current demands on your time allow. It would be especially helpful if, throughout your account, you were to bear in mind <u>how your early experience with the CTP differed from later on.</u> For example, there may have been

a temptation at first to revert to structural explanations; with more experience, perhaps you found you were more able to exploit opportunities for 'negotiation'; and so on.

The greater the detail and the more examples and anecdotes you can recall, the richer the basis for understanding you provide.

This is, of course, a time-consuming undertaking for you, but I regard such accounts as potentially the most fruitful source of all for an appreciation of an innovative teaching methodology. I hope, therefore, that you will respond fully.

Once again, thankyou.

Teachers' accounts: Some Guidelines

1. KNOWLEDGE

In your perception, what was the CTP about?

a) Principles and methodology.

b) Differences between the CTP and other ways of teaching.

c) Changes in role relationships between you and your pupils.

d) Explicit attention to language.

e) Problems that typical teachers might have with the CTP.

f) Other ways of teaching that might have improved the effect of the CTP.

g) Day-to-day requirements of teaching on the CTP.

h) Effects of the CTP on pupils.

2. ACQUIRING INFORMATION

What steps did you take to find out more about the CTP?

a) Did you actively seek information about the project or wait for it to be presented to you?

b) Did you ask other people's opinions about the CTP?

c) Did you attend discussion seminars or receive training related to the CTP?

d) Did you think that there might be particular problems implementing the CTP in 'your' school? If so, did you communicate these doubts to Dr. Prabhu?

e) Did you feel that you ever reached a stage when you no longer required more information about the CTP?

f) Did you try to find out about other ways of teaching that might have enhanced pupil learning?

g) Did you work with other teachers to produce CTP materials?

h) Did you consider making major changes in the CTP approach?

3. SHARING

In what ways, if any, did you share problems, ideas, and materials with other teachers?

a) When you talked about the CTP to other teachers, or to Dr. Prabhu, what sorts of things did you discuss?

Discipline and class control? When you would get more materials? Your current use of the CTP? Your ideas for modifying the CTP? The possibilities of producing your own materials? Working with others to produce materials? Possible alternatives to the CTP?

4. ASSESSING

In what ways did you assess the strengths and weaknesses of the CTP?

a) Did you analyse the CTP before you participated in it? So that you would understand what to do? So that you would be able to judge its likely effects?

b) Did you ever examine your own use of the CTP in terms of classroom management? Or how long you took over pre-task, task, and feedback?

c) Were you always interested in how much pupils were learning with the CTP? Did you try to introduce improvements to influence pupil learning? Elaborate.

d) Did you notice or consciously seek evidence of the merits of the CTP?

5. PLANNING

Did you make short-term or long-term plans about what you would do in CTP lessons?

a) How far ahead did you plan your lessons? Immediate use for the coming week? Several weeks?

b) Did you think that there were steps that needed to be taken to accommodate long-term issues (e.g. external

examinations)? If so, what steps did you take?

c) Did you ever plan activities that would slightly or even considerably modify the CTP?

6. STATUS REPORTING

What was your personal stand in relation to the CTP?

a) Did you feel that a lot of your time was taken up with obtaining materials and working out how you would use them? Or did you feel that your use of the CTP went satisfactorily with few, if any, problems?

b) Did you develop your own materials? Were they consciously different in any way from the materials that were given to you?

7. PERFORMING

What actually happened in the classroom?

a) Did you notice moments when you allowed opportunities to negotiate to slip by? Or when you exploited them? Elaborate.

b) How long could you continue a series of lessons on the same theme before you sensed a need to do something fresh?

c) Did you find that you got into a routine in which you could do all that the CTP required without having to change your teaching very much?

d) Did you experiment with combinations of the CTP and structural teaching (or any other approach) to improve

pupil learning?

e) In what ways, if any, did you adapt the CTP to suit your particular pupils in your particular school?

f) Did you notice if particular pupils benefitted most from the CTP approach? Was it your impression that there were other types who would have been better off being given grammatical rules? Specify.

g) Do you recall moments when pupils seemed to have learnt something that you had not specifically taught?

h) What difficulties did you perceive in making the transition from structural to CTP teaching?

i) Did you ever divide classes into groups for any reason?

j) Did pupils ever talk to each other in English? Or only when talking to the teacher? Did you ever talk to the pupils in Kannada or Tamil? For purposes of discipline? Classroom management? To explain something particularly difficult?

k) How did you deal with grammatical errors? Lexical errors? What did you do if the answer to a problem was wrong but the grammar was right?

5.3.2 Teachers' personal details

Teachers were requested to complete short forms giving personal details. The form is presented here:

Personal Details

Name

Age

Qualifications (at the time of teaching on the CTP)

No. of years ELT experience (prior to teaching on the CTP)

Were you a full-time teacher at the school(s) where you taught on the CTP?

If not, what was your full-time occupation at that time?

How long did you teach on the CTP?

When? From: To:

At which school(s) did you teach on the CTP?

How did you come to be involved with the CTP?

The rationale for collecting this biodata is that variables such as experience, duration of involvement with the project, age, and whether or not the teacher was a regular teacher at the school or an 'outside expert', might account for differences in Levels of Implementation. In other words, it was anticipated that these variables might provide explanations of our

findings. This would be particularly useful information for other researchers or teachers who may require to know what teacher variables are likely to lead to what kinds of implementation. Analysis of the data is reported in section 5.4.1.3.

5.3.3 Number and nature of teachers' responses

In all, 18 teachers took part in the Bangalore project teaching, apart from the director, Dr. Prabhu. Of these, 16 responded. The 2 who failed to respond had only been associated with the project for an exceedingly short period (Prabhu, personal communication). Of the 16 who did respond, one account had to be discounted as it was very short and totally irrelevant. This account is from Teacher P, and is included in Appendix 7.

It was mentioned above (section 5.3.1) that teachers were not expected to pay the postage costs of returning their accounts (and SoC questionnaires; see chapter 6) and that Prabhu had elected to receive these accounts and forward them from Madras. Only one teacher responded directly to the author; the rest all went through Dr. Prabhu. It might be argued that knowing that the director of the project could see their accounts would have a distorting effect on what would be written. It must be conceded from the outset that this remains a possibility; however, the candid nature of the accounts and the, at times, blatant criticism of the project and of its management, certainly help to dispel such doubts.

Furthermore, the fact that critical accounts were, in fact, forwarded by Dr. Prabhu indicates the openness with which the venture was approached in all quarters, and increases confidence in the reliability of the data (cf. chapter 6, section 6.3.2.3).

The nature of the responses varied quite considerably, as predicted. Some teachers wrote long essays apparently unconstrained by the guidelines. Others wrote accounts that would be elliptical if one did not constantly refer to the guidelines, i.e. they might note that with regard to, for instance, probe c) in Performing, this had not occurred; to know what 'this' is, reference must be made to the guidelines. There were also some teachers who simply annotated the guidelines with yes/no answers or with longer chunks of prose. Some accounts were typed; most were handwritten. As the nature of the responses was so varied, they were transformed to a common format (see section 5.3.4 below).

5.3.4 Transformation of accounts to a common format

For ease and consistency of analysis, the accounts were transformed in the following ways, increasing the uniformity of their presentation. First of all, if an account could be fully comprehended independently of the guidelines, none of the guideline probes or questions were included. Where a segment of an account requires mention of these probes to clarify the meaning, it is

inserted and prefaced by 'Q:' for 'question'; the teacher's response follows and is prefaced by 'A:' for 'answer'. The accounts, including the guideline probes as necessary, are presented on the left half of each page under the heading 'Teacher's account'. On the right half of the page, under the heading 'Comments', we have arranged our evaluation comments to be parallel with the portion of text that indicates that a certain level should be assigned for Knowledge, Assessing, Acquiring Information or Performing. This enables other interested persons to see just what it was that a teacher had said which caused the evaluator to come to the judgments he did, and to find them credible or otherwise. So, for example, teacher E recalls trying to find out what impact the CTP was having on pupils by constructing a test which attempted to be program-fair and administering it to both a CTP group and a group receiving the regular structural program; in the parallel 'Comments', we have noted "a clear example of Assessing Level 3. An attempt at program-fair comparison". This behaviour relates to one of the indices for Assessing Level 3 listed in section 5.2.3, above.

Thus, as can be seen from Appendix 7, a series of evaluator comments and allocations to appropriate Levels of Implementation accompanies each teacher's account, and each account is fully comprehensible without frequent and inconvenient recourse to the guidelines. This conversion to a regular format is a preliminary step to the transfer

of information to profile sheets, the subject of the following section.

5.3.5 Transfer of Information to Profile Sheets

The next stage is to transfer the comments to a CTP Teacher Implementation Profile Sheet (adapted from Loucks, Newlove and Hall, 1975). These Profile Sheets are displayed for each teacher in section 5.4.1.1 below. Each mention in the 'Comments' column of an assignment of a particular Level of Implementation relating to Knowledge, Acquiring Information, Assessing and Performing, is checked onto the appropriate part of the Profile Sheet. Thus, the Profile Sheet for Teacher A is checked once for Knowledge at the Orientation level; twice for Acquiring Information at the Orientation level and once at both the Routine and Renewal levels; twice for Assessing at the Orientation level; once for Performing at the Orientation level and once at the routine level.

It is immediately evident that different parts of each teacher's account can evoke an evaluator response at varying levels. This is because, for example, a teacher who might rate level 3 Assessing for attemting to carry out a program-fair testing study might also mention informal attempts at feedback, which would rate a level 1 allocation. Therefore, if an overall Level of Implementation is to be arrived at, a simple procedure is required to determine what level (where more than one has

been checked) should take precedence.

The most obvious method is simply to tally the frequency of checks for each category and then the frequency across categories. In most cases, frequency turned out to be a sufficient criterion, as frequency correlated highly with the overriding impression gleaned from a reading of each account. However, in some cases, there are an equal number of checks for two levels within a category. In such cases, or whenever the frequency runs counter to the overriding impression, the impression takes precedence.

It can be seen from a glance at the Profile Sheets that the level determined within each of the 4 categories is ringed. In all except 2 cases, if 2 or more categories are ringed at a particular level, that level is also the Overall level (on the right of the Profile Sheet). The exceptions are teachers F and J. In both cases, the evaluator's impression took precedence. Since impression rating was judged necessary, brief reports of total impressions of each account follow the Profile Sheets in section 5.4.1, below.

Adhering to the criteria proposed in the current section, all teachers were assigned a single Overall Level of Implementation.

5.4 Results

5.4.1 Teacher Profile Sheets, Impressions and Overall LIs

In this section, each teacher's Profile Sheet is displayed (beginning on the following page, in section 5.4.1.1); then the evaluators total impressions are recorded (in section 5.4.1.2); after this, the Overall LIs are summarised and discussed (in section 5.4.1.3).

CTP Teacher Implementation Profile Sheet

Teacher Profile A

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1 I	1 II	1 II	1 I	1
Routine	2	2 I	2	2 I	2
Renewal	3	3 I	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile B

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1 |	1	1
Routine	(2 |)	(2 | |)	2	(2 | | | | |)	(2)
Renewal	3	3 | |	(3 | | |)	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile C

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1 1 1	1 1	1 1	1 1 1 1	1
Routine	2	2 1	2	2	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile D

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1 ||||	1 ||	1
Routine	2 |	2 ||	2	2 |||	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile E

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	11	1	111	1	1
Routine	2	211	2	21111111	2
Renewal	3	3	31	31	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile F

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1	1 1 1 1 1 1 1 1 1	1
Routine	2 1	2 1	2	2 1	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile G

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1 (	1	1 | | |	1	1
Routine	2 | |	2 | | | |	2 | |	2 (| | | | | |	2
Renewal	3 | | |	3 |	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile H

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	11111	11	1
Routine	211	21111	2	211111	2
Renewal	31111	311	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile I

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1111	1111	1
Routine	21	21	2	2111	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile J

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	11	11	11111	1	1
Routine	2	2111	2	211111	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile K

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	(1 1)	1	(1 1 1)	(1 1)	(1)
Routine	2	(2 1)	2	2 1	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile L

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1 |	(1 |)	1 /	1
Routine	2 |	2 |||	2	2 ||	2
Renewal	(3 ||)	(3 ||)	3	(3 |||)	(3)
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile - M

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1 |	1 |	1 ||||	1 ||||	1
Routine	2	2 ||	2	2 ||	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile N

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	11	1	1111	11111	1
Routine	2	2111	2	2111	2
Renewal	3	3	3	3	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile 0

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1 ||	1 ||	1
Routine	2|	2 |||	2	2 ||	2
Renewal	3|	3 |||	3|	3 ||	3
No information in account:	NI	NI	NI	NI	

General Comments:

CTP Teacher Implementation Profile Sheet

Teacher Profile P

Level	Knowledge	Acquiring Information	Assessing	Performing	Overall Level
Orientation	1	1	1	1	1
Routine	2	2	2	2	2
Renewal	3	3	3	3	3
No information in account:	(NI)	(NI)	(NI)	(NI)	

General Comments:

5.4.1.2 Total Impressions of each Account

Teacher A

This teacher ignores most of the probes and the account is elliptical and ambiguous even where responses are given. There is no evidence in the account to suggest that this teacher was comfortable in her use of the CTP or that she had a well-developed awareness of its principles and methodology. She was a regular teacher at the school where she taught on the CTP and her biodata reveal that came to be interested in the project after Dr. Prabhu had "adopted" the school for the CTP. What indications there are suggest that Teacher A is at Level 1 (Orientation).

Teacher B

Our overall impression is that this teacher is at the Routine level of implementation. Only in Assessing is he clearly at the Renewal level. With regard to Acquiring Information, the possible indications of level 3 are fairly weak and lack illustrative detail. As far as Performing is concerned, there is an overall sense of discipleship; there is a strong awareness of being in the vanguard of an innovation, and the suggestion of a relatively uncritical stance. A concern for the 'purity' of the 'experiment' motivated this teacher during his association with the CTP, and long after the experiment is over, there persists an unwillingness even to consider

modification.

Teacher C

Teacher C was a regular teacher at the school where he taught on the CTP. According to his biodata, he became involved with the CTP because (as with Teacher A), Prabhu had 'adopted' the school for the project. He seems to regard the project as something of an imposition, and took few steps to find out more about it or how to exploit it. He ignores most of the probes in the guidelines. Those that are answered are answered so tersely as to be largely elliptical or ambiguous. So, in spite of 4 years' association with the project, he seems to be barely oriented to the demands of the CTP principles and methodology. It might be speculated that his headmaster's avowed enthusiasm for the CTP (the headmaster is another of the CTP teachers of the present study) had some bearing on Teacher C's long association with it. All the indications are that level 1 is appropriate.

Teacher D

Overall, it seems fair to place this teacher at the Routine level - level 2. The indications of performing at this level are stronger and more numerous than those at level 1 (especially in relation to the production of materials and the treatment of learner error). Teacher D was involved with the CTP for 8 months from the inception of the project.

Teacher E

This teacher is a regular teacher at the school where she taught on the CTP who was recommended to the project team by "Madras Centre". She seems to have been well able to cope with the demands of the innovation. Although Knowledge of the CTP seems preliminary, she reports classroom procedures that are entirely consistent with the CTP pedagogic perceptions. Assessing is well-developed, and even shows an attempt at program-fair comparison. Overall level 2 seems appropriate.

Teacher F

This teacher reveals that he took part in the project reluctantly, and now sees his participation as a mistake (as stated in his biodata). It is clear that he was unwilling to adjust his teaching procedures even for the sake of the 'experiment'. Attempts to nudge him in desired directions are continually viewed as unwelcome. Although he has a reasonable grasp of the nature of the CTP, he never seriously tried to orient himself to the demands of the project. Overall level 1 seems the only possibility.

Teacher G

Teacher G was extremely difficult to categorise. She has been placed at level 2, but it is suspected that there is an equally strong case for level 3 (Renewal). The difficulty is that the various statement s she makes about the standardisation of the CTP seem to be inconsistent. At times she appears to be in favour of

flexible use of the innovation; at others, to reject any adjustment atall. The eventual decision to place her at level 2 derives from an overall sense that indeed there was a set of pedagogic rules to be adhered to, as her statement "no activities were planned to modify the CTP atall" testifies. Nevertheless, it is worth pointing out that allocation of an LI in this instance was precarious.

Teacher H

This teacher senses limitations in the CTP's exclusion of a focus on form. Importantly, and unlike teacher F, this is not from a reluctance to make the transition from structural to CTP teaching but the mature judgment of a teacher who has been teaching on the CTP for a prolonged period (3 years). His Assessing, though casual, leads him to believe that CTP students' learning could be speeded up with some attention to form. His independent awareness, however, is never translated to the classroom for fear of contaminating the experiment. This suggests that the CTP entails fixed behaviours, for example, in relation to the desirability of group work (which he wishes to introduce but does not do so, largely because of this perception of an orthodoxy). Overall, teacher H seems to have an awareness of possible modifications of the CTP, but a strict sense of what must be adopted in practice. Level 2 appears appropriate.

Teacher I

Teacher I is the only one to clearly distinguish between early and late implementation. This means that

although there is frequent mention of difficulty, suggesting the Orientation level, this is because it is the orientation which is being described (with regard to early use). This is especially so in Performing. Behaviour relating to late implementation was used for allocation of Overall LI 2. Incidentally, he does group work without the scruples of teacher H.

Teacher J

There is a clear sense in this teacher's account that the CTP is seen as a set of prescribed rules of behaviour that he tries to adhere to. The impression that comments like "I did not deviate" and "followed the model" give is one of discipleship. The Performing indications suggest that while his Knowledge of the CTP may have been slight, he was able to carry out what he saw as its bidding. Overall level 2 seems fair. Curiously, he considers that the degree of learner-learner interaction "is very great with least effort in the CTP, and it is not so in other methods".

Teacher K

There is a fairly clear sense of wishing to do as little as possible, of going through ready-made tasks with little sense of planning. Also, teacher K tends to emphasise, not his own implementation but the potential for the CTP to be implemented on a grand scale. Overall level 1 seems appropriate.

Teacher L

Our impression is that teacher L's account is a clear example of an Overall level 3. A high degree of understanding of the nature of the CTP is manifested, and this understanding is used to introduce adaptations and modifications. She devises tasks that are thought likely to appeal to the "rowdy" children in her class, and feels free to engage their emotions (where the CTP would prefer to concentrate on mind-engagement), the idea being, presumably, that the confinement to cognitive tasks is artificial and unnecessarily narrow in the kind of language that is likely to flow from it. This teacher takes the CTP idea, explores it, and puts into practice modifications that she judges will benefit learners. A clear level 3 implementor.

Teacher M

The main impression gathered from this teacher's account is that she set about her CTP teaching with great enthusiasm, but at the same time was not taking any chances - she notes that she drilled the weak students separately out of class. Her Knowledge of the CTP is apparently quite sparse and the Performing indicators are primarily at the Orientation level. She was already working at the school when the CTP team arrived there. Level 1 seems reasonable.

Teacher N

This teacher did not give much information. However, the evidence, such as it is, suggests that there was some confusion both in Knowledge and Performing;

overall, we cannot assume that this teacher was implementing the CTP at anything beyond the Orientation level.

Teacher O

The overriding impression from this account is that this is a teacher who is very committed and able to think independently about the CTP. She introduces her own ideas to extend prevailing CTP practice (e.g. she shows considerable initiative in breaking with the strong convention of a teacher-fronted classroom, at least to the extent of having student-class interactions; she does this to encourage more productive language than she perceives many CTP classes permit). She has also discussed the CTP with many non-CTP colleagues and delivered papers at the annual review seminars in Bangalore. She did not mention it in her account, and it does not affect her LI, but she had begun, in the later stages of the project, to insist that her pupils answer verbal questions in complete sentences, simply because she thought them capable of it. (This would seem, however, to clash with the CTP's notion of an incidental struggle with language). All in all, level 3 seems quite appropriate.

Teacher P

This teacher apparently now runs his own school. It is difficult to infer more than that from the account, as it is exceedingly brief and impossible to interpret. It

is probably best considered a null effort - No Information (NI) on the Profile Sheet. (The same teacher failed to complete the SoC questionnaire properly; this, too, had to be discounted.)

5.4.1.3 Overall LIs: Summary and Dicussion of Results

In this section, we look at the overall distribution of LIs among teachers and seek to explain it. (As only teacher I distinguished between early and late implementation, the developmental aspect of the inquiry could not be pursued).

Each teacher's LI is listed in Table 5.2. The summary of this list is contained in Table 5.3. It can be seen from Table 3 that 40% of the CTP teachers were rated at Level 1, 47% at Level 2, and 13% at Level 3. Lumping together levels 2 and 3, it is evident that 60% of the teachers implemented the project at least to the Routine level, while 40% failed to orient themselves fully to the demands of the innovation.

Table 5.2

Overall LIs for 16 CTP teachers

Teacher	LI
A	1
B	2
C	1
D	2
E	2
F	1
G	2
H	2
I	2
J	2
K	1
L	3
M	1
N	1
O	3
P	NI

Table 5.3

Summary of LI allocations

Level	N	%
1 (ORIENTATION)	6	40
2 (ROUTINE)	7	47
3 (RENEWAL)	2	13
NO INFORMATION	1	

The question then is what variables best account for the observed discrepancies in implementation. To try to answer this, the biodata collected from the teachers were analysed. This anaysis will now be discussed with constant reference to the 9 relevant tables located in Appendix 8.

The biodata relate to 8 variables that might have had a bearing on teacher behaviour. These are (1) age, (2) ELT experience, (3) duration of CTP teaching (4) the period when a teacher taught on the CTP (i.e. before or after 1982), (5) school, (6) the teacher's qualifications, (7) the teacher's occupation (if not a regular school teacher) and (8) whether the teacher was a regular teacher at the school (RT) or a non-regular teacher (NRT). We shall examine each of these variables in turn. (All of the data are summarised in Tables 1, 2, and 3 in

Appendix 8, and then treated variable by variable in Tables 4 - 9 in Appendix 8.)

There is no educational or applied linguistic research known to the author which indicates that age is an important variable in the implementation of innovations. Therefore, there was no anticipation that it would have a bearing in the present study, which indeed appears to be the case. It can be seen from Table 1 (in Appendix 8) that the age of teachers when they began using the CTP ranged from 26 to 48 years (excluding Teacher P). Table 4 (in Appendix 8) shows that the average age overall was 36, that the average age for teachers rated at level 1 was 37, also 37 for level 2, and 29 for level 3. As there were only 2 teachers at level 3, the interest is more on levels 1 and 2, where clearly, there is no difference. We may conclude reasonably confidently that age was probably not a variable that could account for different degrees of implementation in the present study.

The range of ELT experience prior to the project ranged from 5 to 26 years (Table 1, in Appendix 8). It may be seen from Table 5 (in Appendix 8) that overall, the mean length of experience was 14 years, that for teachers rated at level 1, it was 15 years, for level 2, also 15 years, and for level 3, 7 years. Once again, the level 3 difference could easily be due to chance as it involves only 2 observations (and we already know that these teachers were younger). There is no difference

between levels 1 and 2, leading us to believe that length of prior ELT experience was probably not a variable contributing to the present study. (None of the teachers was a beginning teacher.)

Table 6 (in Appendix 8) indicates that duration of CTP teaching is unlikely to have had a bearing on the level of implementation. The overall average is 21 months, at level 1, 20 months, level 2, 22 months, and level 3, 24 months. It might have been expected that the longer teachers were associated with the project, the higher their level of implementation would be. However, we have noted in section 5.4.1, above, that teacher C, who taught on the project for 48 months, was rated at level 1; it was speculated that the reason for this teacher's long participation could stem from the enthusiasm of the headmaster (who was one of the earliest CTP teachers).

The educational literature suggests that a relationship between time and implementation exists. Hall and Rutherford found that

> Change is not a discrete event that occurs at some point in time, but a process that occurs over time. The more complex the innovation, the longer it will take to arrive at a point where the innovation is used routinely. (1983, p.2.)

Loucks and Melle came to the same conclusion: "implementation requires a significant time investment. It took 3 years before most of the teachers reached at least a routine level of use" (1981, p.28).

Time may be a necessary condition (although we do not know how much time), but in the case of the CTP it was not a sufficient condition. We shall see when we come to discuss the distinction between regular and non-regular teachers that a sufficient condition may be a sense of 'ownership'.

Given that the CTP evolved through a process of trial and error, it seemed possible that the present study was rating teachers who had taught in the early months of the project's life on criteria that only emerged later on. Although in drawing up the criteria, careful attention was paid to the early literature on the CTP (i.e. the RIE Newsletters and Bulletins), this still seemed to be a variable worth exploring.

The strategy was to try to find a time by which the project principles and methodology were fairly settled, that is, when they were unlikely to undergo further substantial revision. We wrote to Prabhu, asking him to consider this. He replied (Prabhu, 1986, personal communication) that "basic principles remained constant throughout the five years" and that the pattern of classroom activity was "most settled in the last two years (June 1982 to April 1984)". He further refined this, adding that teaching at a post-initial level was "well-settled" by around December 1980, and teaching at the zero level by around December 1981 (for a more

complete rendering of Prabhu's reply, see chapter 7, section 7.3.3). The teachers' accounts do not tell us whether the classes they taught were at the zero level or not, so the refinement offered by Prabhu could not be exploited in our analysis. However, it was possible to separate teachers who had taught pre-June 1982 and those who taught post-June 1982.

Table 7 (in Appendix 8) shows the number of teachers who taught pre- and post-June 1982 and their LI rating. Of those rated at level 1, 2 were 'pre' and 5 were 'post'; of those rated at level 2, 5 were 'pre' and 3 were 'post'; and of those rated at level 3, 2 were 'pre' and 1 was 'post'. (These figures include twice each the 3 teachers who taught both 'pre' and 'post'.) It is not clear from this array of figures whether there is large difference. Some statistical analysis, it was thought, would help to indicate the extent of the difference.

A 2 x 3 chi-square test of significance was considered, but some of the expected frequencies were less than 5. As Siegel (1956, p.110), Robson (1973, p.88), Hatch and Farhady (1982, p.170), and Woods, Fletcher and Hughes (1986, p. 144) note, a rule of thumb is that chi-square is inappropriate when any expected frequency falls below 5. (Although they stress that this does not refer to observed frequencies, Henning 1986, p.706, takes a different view; Guilford and Fruchter 1978, p.203, also stress that it is expected, not observed frequencies that are relevant, but they

consider that the number may fall as low as 2 before chi-square and Yates' correction are inapplicable). The general view, then, is that 5 is the lowest allowable expected frequency; for lower frequencies, cells can be lumped together to increase the frequency, as long as this does not eliminate useful distinctions (in our case combining levels 2 and 3 still did not raise the expected frequencies sufficiently); or chi-square can be computed and reported with appropriate caveats; or the probability has to be computed directly, using the Fisher exact probability test. We chose the latter option.

Combining levels 2 and 3 (we were mainly interested in the distinction between adequate and inadequate implementation, and level 3 involves only 2 teachers) so that we had a 2 x 2 table, Fishers exact test was applied. The result was $p = 0.1425$. Although this is well short of the .05 level of significance, it is clear that there is a marked relationship between pre- and post-June 1982 CTP teaching and LI.

If this relationship had been in the expected direction, it would have been reasonable to stop there and note the relative strength of the variable. However, the LI rating was higher for 'pre' than for 'post' teachers. As we will see in the discussion of the relationship between LIs and RT and NRT teachers below, there is another factor which helps to explain this finding.

The variable 'school' can be dismissed quickly. As can be seen from Table 8 (in Appendix 8), 7 schools are represented among the 15 teachers, so there are far too few observations per school to make any inference.

Qualifications and occupation may be treated together as the former tends to have a bearing on the latter. It may seem strange to investigate occupation in a study which is only concerned with teachers. The position was not so simple, however; many of those who took part in the project were not regular teachers in the schools concerned but were drafted in from higher echelons of the profession. The obvious distinction, then, is between regular teachers (RTs) and non-regular teachers (NRTs).

Tables 2 and 3 (in Appendix 8) list the qualifications and occupations of RTs and NRTs, respectively. All of the regular teachers, obviously, reported their occupation as 'teacher' (Table 2), and apart from one who had obtained a local (Indian) bachelor's degree, they all had only teaching diplomas. A glance at Table 3 reveals that the NRTs were far more highly qualified and employed at higher levels of the educational hierarchy. All NRTs had obtained at the very least a master's degree, and 2 had acquired Ph.Ds. 6 out of 11 (7 out of 12 if we include teacher P; 8 out of 13 if Prabhu is included) had gained ELT-related qualifications in U.K. universities. Most NRTs were employed at teacher training institutions

(mainly the Regional Institute of English, Bangalore), though university and British Council staff were also represented. The overwhelming picture is that the NRTs were a quite different population from the RTs.

The relationship between LIs and RTs and NRTs is presented in Table 9 (in Appendix 8). It can be seen that 75% of the RTs (i.e. 3 out of 4) were still struggling to come to terms with the demands of the project and were rated at level 1, while 73% (8 out of 11) of the NRTs were rated at least at the Routine level (level 2). As with the pre- and post-June 1982 frequencies, it is difficult to see how significant this finding is, so Fisher's exact probability test was applied. (Chi-square was ruled out because some of the expected frequencies were below 5).

Fisher's exact test yielded a probability of $p = 0.1319$. Although not significant at the .05 level, which suggests that it cannot be argued strongly that there is a difference, there is still a considerable degree of association between LI and the RT/NRT distinction.

Our confidence in the validity of this association is increased if we consider (i) the pre/post 1982 finding, reported above in this section, (ii) the discrepancies in qualifications and occupations between RTs and NRTs, also reported above in this section, (iii) the reasons given by CTP staff for joining the project and (iv) the sense of 'ownership' of the project

evidenced in some teachers' accounts (and the corresponding importance of this variable reported in the educational literature).

With regard to (i), the pre/post 1982 finding that there was a marked (though, again it should be stressed, not significant) difference between the LI of 'pre' and 'post' teachers and that the 'pre' teachers were rated more highly was counterintuitive. It would have been more plausible if the direction of the difference had favoured those who taught after the project had become comparatively stable in its methodology. This perplexing finding can probably be best explained by the fact that 4 out of 6 of the 'post' teachers were also RTs. Given the markedly lower LI rating obtained by RTs, this would serve to weight the 'post' population at a lower level. Thus the pre/post 1982 and the RT/NRT findings triangulate.

As for (ii), the fact that NRTs were far more highly qualified than NRTs would not necessarily entail that they were better teachers. However, the level of English of the RTs (and of many typical teachers in South India) was also considerably less developed. Furthermore, there was no indication during my visit to the project or in their accounts that RTs had done anything but the traditional structure-based teaching previously, whereas the NRTs, as teacher-trainers and often fresh from U.K. ELT courses, were possibly more open to ideas.

Turning now to (iii) and (iv), it can be seen from the answers given to the question in the 'personal details' biodata form (Appendix 9) and from the accounts (Appendix 7) that CTP staff were recruited to the project in different ways. The RTs became CTP teachers for the following reasons: two because "Dr. Prabhu ... has adopted this school for CTP", one because she became interested while the project was running at her school, and one because she was recommended to the project team by a teacher-training centre.

It is quite another story with the NRTs. Many of them were aware of being part of a close-knit group who were responsible for developing the approach rather than just carrying it out (though one NRT expressed reluctance to join the project team; see teacher F, in Appendix 9). Teacher J remarks that "Dr. N.S. Prabhu took me into his confidence and helped me join the project team" (Appendix 9); he speaks of "our fold" and notes that "we were the beginners" (Appendix 7). Teacher G says "I worked on the project in its inception ... so we were in a way creating/producing" (Appendix 5B). Teacher O remembers a sense of "for the first time" being "part of a team" (Appendix 7). Teacher H, who was "committed to the CTP" (Appendix 7), refers to "those of us closely involved in the inception and growth" (Appendix 7), uses the term "pioneer teachers" (Appendix 7), and recalls that "it was a privilege to work with a team of experts" (Appendix 9). Teacher H also comments that "we - the members of the CTP

- constructed the method bit by bit" (Appendix 7). To the NRTs, the CTP was theirs.

The notion of possession as a crucial factor in the implementation process has considerable backing in the educational research literature. Martin and Saif advocate what they call "teacher owned" (1984, p.4.) curricula, and add that for classroom change to be effected, teachers must "sense both a personal and professional stake in the proposed change" (p.4.). Williams and Hull (1968) and Hall and George (1979) stress the importance to the individual of being among the first to adopt an innovation. Crandall, Bauchner, Loucks and Schmidt (1982), in a study of dissemination efforts supporting school improvement, affirm that "commitment to, or 'ownership' of, the innovation is vital to succesful implementation" (p.6.), a point also made by McLaughlin and Berman (1975).

The pre/post-1982 finding, the superior qualifications and ELT background of the NRTs, the process of recruitment to the project and the sense of ownership expressed by NRTs all combine to indicate that the best interpretation of our LI data is that level of implementation depended on whether or not the teacher was a regular teacher in a school where the project operated.

5.4.2 Typical teachers' difficulties with the CTP

That the regular teachers in our study found

difficulty implementing the CTP is not surprising if we consider that they themselves thought (as we shall see below) that the innovation would raise major problems for typical teachers in South Indian schools. In this section we put the views of both RTs and NRTs concerning the difficulties facing typical teachers. This is worth knowing because one of the potential uses of the present evaluation is that interested practitioners (administrators or teachers) may wish to consider implementing either a similar approach under similar conditions, a similar approach under dissimilar conditions, or a dissimilar approach under similar conditions. In all three cases, information provided by the 'front line' users would be relevant. That is to say, we are interested not only in arguing that typical teachers (the RTs) in the Bangalore project experienced certain problems with regard to implementation, but also in facilitating extrapolation by interested parties to other typical teachers beyond the CTP and beyond the project schools. According to Guba (1981), extrapolation is the appropriate aspect of generalisability for a naturalistic study (such as the present chapter reports); for Cronbach (1982), the concept of extrapolation is central to program evaluation, and he gives prominence to it in his concept of external validity (cf. chapter 2, section 2.4).

Examining the teachers' accounts (Appendix 7), the difficulties facing typical teachers receives frequent

and detailed mention. Clearly it is a salient issue for both RTs and NRTs. The same problems tend to recur through most of the accounts. They are (i) excessive demands made on time, (ii) discipline, (iii) insufficient command of English and (iv) a tendency to revert to structure-based teaching. We shall look at each of these issues in turn.

Time is seen as the most prominent impediment of all to typical teachers' implementation of the CTP. Teacher C states simply that "CTP consumes a lot of my time", but does not elaborate or extend the statement to other typical teachers. Teachers A, E and K specify that the time is consumed by preparing materials and correcting students' work; teacher K observes that CTP materials would "have to be given to the teacher, who cannot be expected to take on the additional role of materials writer. The Indian teachers are a hard-worked lot and they are hard-pressed for time."

Teacher L notes that not only is materials production onerous, but also that what exacerbates the problem is the huge numbers of students in a class:

> what might be genuinely problematic is the time requirement of the CTP. Preparing for a task and providing feedback on tasks are time-consuming in themselves. Given our large classes and small number of teachers, this can be seen as a practical hurdle...

Some teachers see discipline as a potential hazard. As teacher H speculates, "typical teachers might have problems in coping with the class, i.e. judging how much

freedom learners can have". This is corroborated by teacher M's acknowledgement that "there is no discipline, control over the class". As to why there should be less discipline, the most plausible explanation seems to be the slightly changed role relationship between learner and teacher (slightly, because in the CTP, the lesson is still teacher-fronted) and because of learner's expectations. Teacher H comments that

> students had made their perception that an English class should have all the factors of a structural class such as drilling, teacher providing all help, etc., and so it took a long time to shake them off their image of an English class.

As for command of English, teacher I takes the view that "typical teachers might have problems with their own English which is in most cases insufficient". Teacher E (an RT) judges that they need "the command on the language which is very important for the success of this approach". Teacher H ventures the opinion that they will have difficulty in "keeping the class going i.e. conversing with the learners in English". Quite clearly, the project teachers think that the CTP requires greater fluency in English than can be expected from most school teachers in South India, a point also made by Davies: "there is, in cases of inadequate proficiency, a tendency to stick close to a 'script'" (1983, p.13).

It is further perceived by many project teachers that typical teachers will tend to revert to structure-based teaching. This is not mentioned by RTs (although

teacher M does report giving slow learners separate lessons in grammar). NRTs acknowledge that they themselves would often be tempted to teach structures, e.g. teachers F, D, H, and I. Teacher D draws the moral: "the structural hangover was evident in some of my lessons ... My point is that for a typical teacher it may be difficult to meet the demands of the CTP." Teacher B thinks that most teachers "trained in a structural methodology have at least in the beginning a tendency to implicitly or explicitly draw the attention of the pupils to the form of the language." Teacher M asserts that teachers would be concerned that the learners "do not know grammar". Presumably with this in mind, teacher J singles out the problem many teachers would have in making "the slow learners confident."

Brumfit, who twice visited the project, commented that

> the methods of the teachers, especially when teachers other than Prabhu or his closest associates have been teaching, have tended to revert to specific teaching of language items (1984, p.238).

There is one further area that project teachers thought would present problems, but it is not easy to define. It appears to verge on a belief that most teachers are not intelligent enough, but it may simply be that it is felt that they would have difficulty in adapting. Teacher L, discussing the difficulties of typical teachers states that

> responding to the task in the pre-task phase calls for absolute familiarity with the task and its possibilities, quick thinking, calculating, inferencing, etc.

and notes that this is extremely difficult compared with what she calls the "soft option" of structure-based teaching. Teacher H puts forward the view that

> another thing which may be common in most of the classes of typical teachers is that they miss opportunities which could have been exploited for learner-participation/negotiation.

Whether it is perceived that these deficiencies are due to teachers' poor fluency, adaptability or intelligence is difficult to ascertain. Nevertheless, there is clearly a further area of potential difficulty.

To sum up, project teachers (both RTs and NRTs) regard the CTP as potentially difficult for typical teachers to implement because of demands on time, a slackening of discipline, poor command of English and a tendency to revert to the more familiar structure-based approach.

5.4.3 Discrepancies between the LI concept and Prabhu's view of implementation

Quite what the CTP intended to achieve has not been easy to determine. Davies (1983) says:

> the problem that I found quite unresolved is whether the project is regarded (by its members) as an experiment, or a literacy/reading drive, or a mission, or a teaching development intended for teachers' enjoyment. (p.11).

Brumfit (1984), too, talks of "uncertainty about the status of the project. It started out as an experiment,

but it soon acquired a momentum of its own as a result of the interest it created." (p.238).

Prabhu mentions that the project "is an experiment", by which he means "a searching exercise, not a selling one; an attempt at self assurance, not at persuading others" (RIE Newsletter, 1979, Vol.1. No.2. p.21). He disapproves of lists of 'dos' and 'don'ts' in program dissemination, stating that "it is because we don't want that to happen that we are putting so much stress on innovation and how it is different from propagation" (RIE Bulletin 1980, Vol.4. No.1. p.80). However, in a report of a seminar reviewing the progress of the project, one participant objects that the excerpt from a CTP lesson under discussion could have been taken from a structural lesson. Prabhu responds that "in that case, our burden of retraining is likely to be reduced" (RIE Bulletin 1980, Vol.4. No.1. p.50).

If no propagation is anticipated, it is perplexing that 'retraining' should be entertained. This sense of ambiguity seems to have conveyed itself to the CTP teachers. Teacher L remembers that

> I, and most of my colleagues at the RIE [most of the CTP teachers were recruited from the Regional Institute of English], saw the CTP as an alternate model offered for dissemination and not as an experimental project. This radical prospect must have been a threat to most people. If the project group had got across to the community right at the beginning that this was only an experiment, I feel the climate might have been better for the CTP. (Appendix 7).

More recently (and as we have briefly mentioned in

section 5.2.2 above) Prabhu (1987, p.103) has stated that his view of "the project - and of pedagogic innovation generally" is not that teachers carefully conform to prescribed practices (an approach widely referred to in the educational literature as 'fidelity', e.g. Loucks, 1983; Fullan and Pomfret, 1977); instead they

> (1) operate with a sense of plausibility about whatever procedures they adopt and (2) each teacher's sense of plausibility is as 'alive' or active and hence open to development and change, as it can be. (Prabhu, forthcoming).

It is no criticism of the quality of these deliberations to point out that on the whole, the CTP teachers' accounts fail to reflect an openness to development and change.

On the contrary, there is a clear sense of trying to follow an approach faithfully. LI level 3 was included in the present investigation primarily as a means of taking into account precisely the kind of independence Prabhu approves of, and only two teachers warranted this allocation. For most CTP teachers, strict guidelines were to be adhered to.

Teacher B says "I always guarded myself against trying to improve the method since that would vitiate the validity of the experiment." In the same vein, teacher D asserts "I think I did not deviate from the principles and guidelines set by Dr. Prabhu." Teacher H judges that "as an experimental project, CTP remained as much uncorrupted as possible in my hands." Teacher N claimed

not to have adapted anything: "CTP was used just as it is." Perhaps the most trenchant comments come from teacher F, who states that "accommodating other ways of teaching was never allowed lest it should weaken/alter the CTP philosophy" and that "any modifications beyond the approved framework were shot down"; these comments are accompanied by an example in which the teacher was stopped by a CTP observer for drawing attention to the difference between 'since' and 'for'.

With regard to group work, teacher H took the view that "more groupwork might improve the CTP", but used it only very occasionally "because the CTP has reservations about the groupwork technique in the classes." (These reservations are discussed in Chapter 3 along with the reactions of Brumfit 1984; and Howatt 1984, and others, to this particularly controversial aspect of the CTP). Teacher O, however, (rated at level 3), consciously produced task-types that called for one student to give instructions for another to draw figures on the black-board. As she explains, CTP tasks, with their lack of learner-learner interaction, though they were cognitively challenging, "linguistically did not make adequate demands on the learners' productive abilities."

Another controversial feature of the project (also discussed in Chapter 3) has been its insistence on purely cognitive tasks and the absence of tasks involving a wider range of rhetorical functions (e.g. affect) (Johnson, 1982; Davies, 1983; Brumfit, 1984; and Barnes,

1982). Again, teachers observed the CTP 'rules'. Teacher J produced his own materials but "followed the model". Similarly, teacher B recalls: "we developed materials of our own but they were not different in any way from the given materials." Teacher E, too:

> Regarding the development of my own materials, they were not different from those supplied to me. I had the model tasks. On the basis of it, I easily prepared the tasks which I needed.

Prabhu (1987, p.52) contends that it would be:

> wrong to imagine that task-based teaching involves treating learners as mere reasoning machines, and it was not the project's experience that reasoning-gap activity was 'dull' for learners.

Teacher L, however, has a different perspective:

> Tasks that are varied, that involved the affective experience of learners and teachers to a greater extent would have reduced the monotony of tasks and made some of us more at ease with the task.

Teacher L (the other teacher rated at level 3) was the only one who consciously devised tasks that did not conform to the cognitive 'model'; "they were different in terms of including tasks involving more affective abilities of learners."

What emerges from an examination of the accounts is that the majority of teachers look back upon the project as a period when conformity to prespecified procedures was required. This contrasts with the notion of the project involving teachers experimenting and modifying according to their individual sense of plausibility, without having to adhere to an imposed set of

methodological guidelines. If this notion had been more widely received, we may have witnessed more of the initiative shown by teachers L and O. In the event, there appears to have been more of a constraining effect than may have been intended.

5.5 Summary and Conclusions

In this section, the ways in which the CTP was implemented by individual teachers were explored. First of all, a rationale was offered for the study. Secondly, an implementation measure was devised and operationally defined, to analyse a database consisting of historical narratives written by 16 CTP teachers recalling their experience while associated with the project. The results of the analysis indicated that 6 teachers remained at the Orientation level of implementation, that is, they never fully came to terms with the demands of the CTP; that 7 teachers were at the Routine level of implementation, which is to say that they were operating with the innovation comfortably and according to stated CTP perceptions; and that only 2 were at the Renewal level of implementation, or consciously seeking ways of improving the CTP in order to benefit learners.

It was considered that the best interpretation of all the evidence was that differences between Orientation and Routine (and Renewal) levels depended on the distinction between regular teachers (RTs) and non-

regular teachers (NRTs). RTs tended to implement at a lower level than NRTs. It was argued that this could probably be best explained by the tokens of 'ownership' of the CTP exhibited by NRTs.

In view of our LI findings, and also of project teachers' expressed doubts, it seems reasonable to judge that the CTP would not be readily assimilible by typical teachers in South Indian schools (or, by extension to other schools elsewhere where similar antecedent conditions pertain).

To end on a note of caution, the study reported in the present chapter is a naturalistic one and dependent on interpretation. We have endeavoured to keep the reporting transparent at every stage of the investigation so that interested parties may make their own judgments and extrapolations. This is a minimum condition for what aspires to be a disciplined inquiry.

CHAPTER 6

STAGES OF CONCERN OF INDIVIDUAL TEACHERS ABOUT THE CTP

6 Stages of concern of CTP teachers

6.1 Rationale

A major aim of the evaluation of the CTP, especially the segments of the inquiry reported in chapters 5 and 6, is to facilitate extrapolation from the Bangalore project to the implementation of similar projects elsewhere. Chapter 5 considers levels of implementation by individual teachers, while chapter 6 investigates the 'concerns' of the CTP teachers. As Brumfit points out,

> At least from this project [i.e. the CTP], suggestions for adaptation to other circumstances can be derived, and similar projects can be set up - as indeed they have been in several parts of India. (1984, p.238).

Evaluation that serves this purpose can aspire to being usable (cf. chapter 2, section 2.4.2).

A factor which is likely to have a bearing on the way an innovative program is implemented is in the relative intensity of teachers' concerns at different phases of their association with the program. If we could determine which concerns were most salient at different times, it might be possible to make provision for them in any further attempt at implementation of the CTP or projects like it. In other words, if we can gain an understanding of the dynamics of individuals involved in implementing the CTP in Bangalore, then it might affect the management of future implementations. For example, if teachers are mainly concerned with simply coping in the first few weeks, there would be little point in stressing

matters of effectiveness.

6.2 Previous concerns research

Until recently, exploration of teachers' concerns focussed principally on the anxieties experienced by trainee teachers, and not on teachers involved in innovations. Phillips (1932), for example, looked into the difficulties faced by beginning teachers. Travers, Rabinowitz and Nemovicher (1952) found that trainee teachers were concerned about discipline and whether or not they would be popular with pupils, and judged that training colleges were failing to address these issues.

Gabriel (1957), in a large-scale study, compared experienced with inexperienced teachers, and noted that the latter were more concerned about maintaining discipline and receiving criticism from superiors, and gained significantly less satisfaction than experienced teachers from pupils' success ($p = .01$).

Similarly, Fuller (1969) noted distinctions between pre-teaching, early teaching and late teaching phases, the former two phases being characterised by general apprehension and concerns about adequacy to cope in the classroom, while the latter phase is characterised by greater concern with pupil learning.

Fuller's work showed that concerns occur in a natural, developmental sequence, moving from a focus on 'self' to a focus on 'impact'.

The body of work outlined in this section, particularly that of Fuller, formed the basis for long-term inquiries carried out by Hall and his colleagues at the Research and Development Center for Teacher Education at Texas University (Austin). This culminated in a thoroughly researched questionnaire which aimed to distinguish between seven Stages of Concern (henceforth SoC) for an individual teacher at any given time. The major difference here was that while earlier research had addressed the problems of trainee teachers, the Texas researchers chose to attend to concerns experienced by teachers involved in innovations. This perspective makes their work relevant to the teachers on the CTP, and is therefore treated in greater detail in the following section (section 6.2.1).

6.2.1 Stages of Concern (SoC) about an innovation.

Many in-house publications have elaborated the SoC concept, but the most complete source of reference is the SoC manual (Hall, George and Rutherford, 1977). This manual describes the various research enterprises that were undertaken to arrive at a questionnaire that could reliably and validly gather information about SoC.

Seven different stages were identified through pilot work, and it appeared that they were developmental insomuch as earlier concerns (about self) had to be lowered before later concerns (about impact) could be heightened. Furthermore, it seemed that the concept was

generic: "The research suggests that this developmental pattern holds for most process and product innovations." (Hall, George and Rutherford, 1977, p.6).

6.2.1.1 Definitions of Stages of Concern

The following definitions are taken from Hall, George and Rutherford (1977, p.7):

0 AWARENESS: Little concern about or involvement with the innovation is indicated.

1 INFORMATIONAL: A general awareness of the innovation and interest in learning more detail about it is indicated. The person seems to be unworried about himself/herself in relation to the innovation. He/she is interested in substantive aspects of the innovation in a selfless manner such as general characteristics, effects, and requirements for use.

2. PERSONAL: Individual is uncertain about the demands of the innovation, his/her inadequacy to meet those demands, and his/her role with the innovation. This includes analysis of his/her role in relation to the reward structure of the organisation, decision making and consideration of potential conflicts with existing structures or personal commitment. Financial or status implications of the program for self and colleagues may also be reflected.

3 MANAGEMENT: Attention is focused on the processes and tasks of using the innovation and the best use of

information and resources. Issues related to efficiency, organizing, managing, scheduling, and time demands are utmost.

4 CONSEQUENCE: Attention focuses on impact of the innovation on students in his/her immediate sphere of influence. The focus is on relevance of the innovation for students, evaluation of student outcomes, including performance and competencies, and changes needed to increase student outcomes.

5 COLLABORATION: The focus is on coordination and cooperation with others regarding use of the innovation.

6 REFOCUSING: The focus is on exploration of more universal benefits from the innovation, including the possibility of major changes or replacement with a more powerful alternative. Individual has definite ideas about alternatives to the proposed or existing form of the innovation.

6.2.1.2 The Stages of Concern Questionnaire (SoCQ)

Once the concept of SoC had been developed, the next step was to devise a means of measuring it reliably and validly. The principal strategy was to construct a questionnaire. 195 items were derived from an initial base of 544, which were sorted into stages through a process of judgment and editing. The 195-item questionnaire was piloted and reduced to 35 items measuring the 7 stages, the final form of the SoCQ (presented in Appendix 10).

The reliability of the SoCQ (internal consistency) was computed using Cronbach alpha, and the stages vary from .64 to .83 for 830 respondents (6 of the 7 coefficients were above .70). Test-retest reliability was also calculated with the Pearson product-moment correlation, and the range was from .65 to .86 for 132 respondents (4 of the correlations being above .80).

Validity was less readily established. Firstly, it was perceived that items purportedly measuring the same stage tended to be responded to similarly (on a scale of 0 to 7 for each item). Secondly, a principal components factor analysis with varimax rotation was interpreted in such a way as to suggest that the stages were actually seven independent constructs (Hall, George and Rutherford, 1977, p.12-13). Thirdly, a relationship was found between SoCQ scores and open-ended statements of concerns; however, the relationship was only moderate (multiple R = .58, not significant at the 5% level) and the sample size was small (N = 27). A fourth study compared the results of 40 teachers who had scored very high or very low on SoC stages 2 or 5 with written accounts of their concerns and responses to detailed descriptions of the Stages of Concern, but inconsistencies were found. A further study also compared ratings of concerns with SoCQ scores and found only moderate correlations. A fifth strategy was for 3 researchers to rate 28 respondents for concerns (from

tapes of Levels of Use interviews, discussed in Chapter 5.2.1) and then compare the ratings with SoCQ scores. Inter-rater reliability ranged from .42 to .85 for separate stages, and significant correlations were found between ratings and SoC score rankings in 6 out of 7 stages. In short, the results of the validity studies do allow a moderate measure of confidence that the SoCQ taps 7 distinct stages, though the evidence is hardly overwhelming.

6.3 Method

In view of the SoC questionnaire's aim to determine concerns of individuals involved in innovations, and bearing in mind the stringent efforts made to establish the reliability and validity of the instrument, the decision was taken to utilise the SoCQ for an investigation into CTP teachers' concerns. Immediately, however, the question arises as to how far an instrument trialled and validated for a population in the U.S.A. would be applicable to teachers in South India. In sections 6.3.1 and 6.3.2 below, certain adaptations to the instrument (to accommodate CTP issues) are discussed and the implications for reliability and validity are stated.

6.3.1 Adaptations of SoCQ for CTP Inquiry

A number of modifications were made to the original SoCQ instrument in order to tailor it to the particular

needs of the present inquiry. The most important change was in its aim to get at recalled concerns rather than current concerns. As in the Levels of Implementation study of chapter 5, the present inquiry is retrospective rather than introspective in nature. Also, certain terminological alterations were made to render the meaning more readily comprehensible (on a purely subjective basis). We will now examine both forms of modification in detail.

With regard to retrospection, the most obvious need was to rewrite the items with past tense markers. Also, in order to retain the developmental aspect of the SoC concept, it was necessary to request respondents to retrospect about more than one period of time. While the original SoCQ would be administered at various times during an innovation, we were limited to one post hoc administration. The original SoCQ, (in its scoring), furthermore, distinguishes between 4 groups of users: (i) non-users, (ii) inexperienced users, (iii) experienced users, and (iv) renewing users. Considering whether this could usefully influence the number of time distinctions teachers could be asked to recall in our study, it seemed likely that teachers would have difficulty making a fourfold distinction. Therefore, so as not to overload the demands made of retrospection, the 4 periods were collapsed to 3 with the following labels: (i) when I was first aware of the CTP, (ii) After a few weeks of using

the CTP and (iii) toward the end of my use of the CTP.

Thus, the SoCQ was amended not only with regard to tense markers but also to include 3 time periods for recall within a single administration.

Several minor adjustments were made in addition, which are primarily terminological. For example, the term 'learner' was used instead of 'student', and 'teachers' instead of 'faculty', as it was my impression that in South India, the replaced terms relate to tertiary level institutions. Additionally, wherever the original SoCQ referred to 'innovation', 'CTP' was preferred. More importantly, with respect to Stage 6 (item number 31), the term 'replace' was deleted, as our inquiry intended to include investigation of modification of the CTP, but not replacement (as elaborated in chapter 5). Quite substantial changes were made to 4 items, one of them in Stage 2, and 3 in stage 3. They are as follows:

Item 8 was changed from: I am concerned about conflict between my interests and my responsibilities, to: I was concerned about keeping all of the learners involved.
Item 25 was changed from: I am concerned about time spent working with non-academic problems related to this innovation, to: I was concerned about discipline and organisation in the classroom.
Item 34 was changed from: Coordination of tasks and people is taking too much of my time, to: Preparing for CTP lessons was taking too much of my time.

In these 3 cases (relating to Stage 3), the concern is about management, and the changes were made to clarify this. However, the modifications were based on subjective judgments about what would be most comprehensible to respondents who were not represented in the samples the original SoCQ was piloted on.

With regard to stage 2, the following alteration was made:

Item 13 was changed from: I would like to know who will make the decisions in the new system, to: I wanted to know whether my involvement with the CTP would help or hinder my career.

Stage 2 is concerned with personal adequacy and status implications, and, again, it seemed that the original item would be obscure to the CTP teachers, especially in that it assumes a decision maker and a new 'system' rather than an approach to teaching. Once again, this judgment is impressionistic.

The revised SoCQ is presented in Appendix 11.

Given both the modifications and the different use of the questionnaire to probe recollected concerns, and given also the fact that the statistics of the original SoCQ were based on samples from a population which has little in common with the CTP teachers, it is perfectly clear that the same levels of reliability and validity of the original SoCQ cannot be assumed in the modified

version.

However, the modifications were sparingly made insofar as very few questions were substantially altered. It might still be reasonable to expect that if the original SoCQ measured what it was intended to measure, then since its character had changed very little, it would continue to measure what it intended to measure. This is, it must be stressed, a rational rather than an empirical appeal, but the alternative of developing an untrialled and unvalidated questionnaire has far less to recommend it.

The major modification resides in the use of the questionnaire for retrospective rather than ongoing self-report, but even here, different aspects of validity may be involved in a trade-off rather than simply loss. As Haynes and Wilson argue, "it might be expected that, while retrospective self-reports (...) may not be as sensitive or valid as ongoing self-reports ... ongoing self-reports may be more reactive" (1979, p.311). That is to say, CTP teachers may respond inaccurately due to faulty recall, but they are less likely to respond inaccurately because of perceptions of behaviour induced by the questionnaire itself. It might also be speculated that demand characteristics (such as responding in ways likely to be interpreted favourably) could be reduced by the distance in time from the event. Again, what validity is lost cannot be quantified, but it is worth noting that there may be gains too.

The only kind of reliability that would be suitable for the SoCQ would be the test-retest method, but in view of the difficulties of administering and collecting the questionnaire from a distance, a second administration was impractical. A measure of internal consistency would be inappropriate because the different scales were not designed to be homogeneous. A split-half technique would be difficult to justify since no arrangement of items would yield even superficially similar halves. The technique adopted in the present study was not a formal, quantitative method, but instead a rule of thumb relating to the similarity of an individual's responses to each of the 7 scales (of 5 items each) which were intended to tap separate stages of concern. This is described in section 6.5.2 below.

6.3.2 Administration and Collection of SoCQ

This section looks at the ways in which the SoCQ was introduced to CTP teachers, how they were requested to fill it in, and what provision was made for distribution and collection. Section 6.3.2.1 presents the 'cover letter'; section 6.3.2.2, an 'introduction to SoCQ for CTP teachers'; and section 6.3.2.3, the distribution and collection procedures.

6.3.2.1 The Cover Letter.

At the same time as CTP teachers were asked to

complete the SoCQ, they were also asked to complete a biodata questionnaire and to write detailed accounts of their experience with the project. Thus, one cover letter, which was sent to each teacher, included mention of all three requests. The cover letter is as follows:

Dear X,

At Edinburgh University, an attempt is currently being made to gather more information about the Communicational Teaching Project. As you participated in this project, you are in a unique position to shed more light on the experience, and therefore I hope that you will feel disposed to complete the enclosed questionnaires and, more importantly, to provide as detailed an account of 'what it was like' as current demands on your time allow.

As you know, the CTP was evaluated early in 1984 by myself and Alan Davies. This evaluation was limited, however, to a testing perspective, and cannot begin to furnish us with the kind of understanding that only the teachers in the project have access to.

May I assure you that any publications citing your responses and observations will mention no names, but will refer instead to 'Teacher A, B, C...etc.' (unless you specify otherwise). Also, copies of any publications citing your data will be sent to you. (Please enclose your address).

In the interests of timeliness, I would ask you to

try and return the account and questionnaires within a month of receiving them to Dr. Prabhu at the above address. (Dr. Prabhu has very helpfully elected to send the enclosures on to me from Madras).

Your co-operation in this 'illuminative' research effort would be greatly appreciated, though I do understand that it is asking a great deal. May I therefore thank you in advance.

Yours sincerely

6.3.2.2 Introduction to SoCQ for CTP Teachers

At first sight, it is perhaps not immediately evident to potential respondents how the questionnaire is to be completed. Thus, an introduction to SoCQ was sent to each teacher, and this is presented below:

Introduction to Stages of Concern Questionnaire

The purpose of this questionnaire is to determine what teachers who have used the CTP approach were concerned about at various stages during their association with the project.

3 stages are of interest:

1. When you were first aware of the CTP.
2. After a few weeks of using the CTP.
3. Toward the end of your use of the CTP.

Please read the questionnaire items and respond to them in the answer sheets provided [included in Appendix

11], according to the instructions below.

For items that are irrelevant to you at a particular stage, please circle "0" on the scale. Other items will represent those concerns you had at different stages, in varying degrees of intensity, "7" representing most concern, "1" representing least concern, as in the following explanations:

0 1 2 3 4 5 (6) 7 This statement was very true of me

0 1 2 3 (4) 5 6 7 This statement was somewhat true of me

0 (1) 2 3 4 5 6 7 This statement was not atall true of me

(0) 1 2 3 4 5 6 7 This statement seems irrelevant to me

Let us take the first item on the questionnaire as an example:

1. I was concerned about learners' attitudes toward the CTP.

If at the first stage 'When I was first aware of the CTP', this statement was not atall true of you, you might circle "1" on the relevant scale; if at the second stage 'After a few weeks of using the CTP', this statement was very true of you, you might circle "6" on the relevant scale; if at the third stage 'Toward the end of my use of the CTP', this statement was only somewhat true of you, you might circle "4" on the relevant scale. Your completed answer sheet would then look like this for question 1:

1	0 (1) 2 3 4 5 6 7	0 1 2 3 4 5 (6) 7	0 1 2 3 (4) 5 6 7

Finally, although you may not find it easy to recall exactly what you felt at the 3 different stages, please try to answer the questions to the best of your ability.

Thankyou for taking the time to complete this task.

6.3.2.3 Distribution and Collection Procedures

The cover letter, introduction to the SoCQ, the SoCQ itself, the biodata questionnaire and the Levels of Implementation materials were sent to Dr. Prabhu in Madras, who had helpfully volunteered to send them on to the CTP teachers. As can be seen from the cover letter (6.3.2.1 above), teachers were asked to send their replies via Dr. Prabhu. It must be acknowledged that the prospect that the director of the project would have access to their responses might have had an influence on the teachers. However, there are two arguments against this possibility. Firstly, even if a respondent had wished to distort his or her responses, it would have been exceedingly difficult to anticipate what set of concerns were likely to produce a favourable impression, as they would probably be unaware of the concerns literature and of the relationship between individual items and hypothesised stages of concern. Secondly, as was seen in chapter 5 (section 5.3.3), the fact that

teachers responded at times critically of the CTP in their narrative accounts suggests that they were not inhibited by the collection procedure. (The procedure was necessary because at least some of the teachers would not be able to, or would not wish to, afford the postage costs. In the event, only one teacher responded to me directly, and this was only because she had neglected to respond until Dr. Prabhu had left Madras).

6.4 Data Analysis

The SoCQ contains 35 items. Each item relates to a particular stage of concern about the CTP. Thus, for the 7 stages, there are 5 items expressing each stage. As can be seen from section 6.3.2.2 above, teachers were asked to rate their concern for each item and each of 3 time periods on a scale from 0 to 7. High numbers indicate intense concern, low numbers mild concern, and zero, either no concern or a sense of irrelevance.

Section 6.4.1 below shows the items arranged according to stage (Table 6.1) and also a quick checklist of each item number and its associated stage (Table 6.2).

6.4.1 Items and Stages

Table 6.1

Items Arranged According to Stages

STAGE 0

3. I did not know anything about the CTP

12. I was not interested in the CTP

21. I was completely occupied with other things

23. Although I did not know anything about the CTP, I was interested in teaching methods in general

30. I was not interested in learning more about the CTP

STAGE 1

6. I had very limited knowledge of the CTP

14. I wanted to discuss the possibility of using the CTP

15. I wanted to know what materials and training would be available if I decided to adopt the CTP

26. I wanted to know what using the CTP would require in the immediate future

35. I wanted to know how the CTP was better than structural teaching

STAGE 2

7. I wanted to know what effect the CTP would have on my professional status

13. I wanted to know whether my involvement with the CTP would help or hinder my career

17. I wanted to know how my teaching was supposed to change

28. I wanted to have more information on time and energy commitments required by the CTP

33. I wanted to know how my role would change when using the CTP

STAGE 3

4. I was concerned about not having enough time to organise myself each day

8. I was concerned about keeping all of the learners involved

16. I was concerned about my inability to manage all that the CTP required

25. I was concerned about discipline and organisation in the classroom

34. Preparing for CTP lessons was taking too much of my time

STAGE 4

1. I was concerned about learners' attitudes towards the CTP

11. I was concerned about how the CTP would affect learners

19. I was concerned about evaluating my impact on learners

24. I wanted to get my learners enthusiastic about their involvement in the CTP

32. I wanted to use feedback from learners to change the CTP

STAGE 5

5. I wanted to help other teachers in their use of the

CTP

10. I wanted to develop working relationships with other teachers using the CTP

18. I wanted to familiarise other teachers or scholls with the benefits of the CTP

27. I wanted to coordinate my effort with others in order to maximise the effects of the CTP

29. I wanted to know how other people were using the CTP

STAGE 6

2. I knew of some other approaches that might have worked better

9. I was concerned about revising my use of the CTP

20. I wanted to revise the CTP's instructional approach

22. I wanted to modify my use of the CTP in view of the effect it was having on my learners

31. I wanted to know how to supplement or enhance the CTP

Table 6.2

Checklist of item numbers and associated stages

Item	SoC	Item	SoC	Item	SoC	Item	SoC
1	4	10	5	19	4	28	2
2	6	11	4	20	6	29	5
3	0	12	0	21	0	30	0
4	3	13	2	22	6	31	6
5	5	14	1	23	0	32	4
6	1	15	1	24	4	33	2
7	2	16	3	25	3	34	3
8	3	17	2	26	1	35	1
9	6	18	5	27	5		

6.4.2 Obtaining Raw Scores and Percentages

Raw scores were arrived at by tallying a teacher's responses to each of the 5 items relating to each stage. Thus, for example, a teacher who responds to Stage 1 items (3, 12, 21, 23, and 30) with ratings of 4, 4, 3, 4, and 3 has a raw score of 18 for that stage.

In order to facilitate interpretation of these raw scores, Hall, George and Rutherford (1977, p.27) convert them into percentiles which are based on the responses of 646 individuals. It has already been noted that the samples used in the Texas studies were from a quite different population from the CTP teachers. Therefore, the percentiles derived from the Texas sample cannot be assumed to be valid for the CTP teachers. However, with a sample of only 15, deriving our own percentiles would be coarse. In view of this, no attempt was made to derive a percentile scale; instead, raw scores were simply converted to percentages.

6.5 Results

A total of 16 teachers responded, though one (Teacher P) failed to attend to all items in the questionnaire, and consequently was dropped from the analysis. (Incidentally, Teacher P also failed to respond adequately to the Levels of Implementation probe; see chapter 5). Altogether, 18 teachers actually took part in the CTP teaching, but the 2 who did not respond atall had

an exceedingly brief association with the project (Prabhu, personal communication, 1986).

Raw scores and percentages for each teacher are presented in Appendix 12, and the percentages are summarised separately in Appendix 13.

In this section, following Hall, George and Rutherford (1977), we examine the results from the (preliminary) perspectives of peak scores (section 6.5.1) and complete profiles (section 6.5.2). Possible interactions between SoCQ scores and RT or NRT status are explored in section 6.5.3; (the Chapter 5 finding that RT/NRT was probably related to Level of Implementation raises the possibility that this factor may also have a bearing on concerns).

6.5.1 Peak Score Interpretation

According to Hall, George and Rutherford (1977, p.29), the simplest form of interpretation of results is the peak score interpretation, in which the most intense concern for each individual teacher is highlighted at the 3 time periods. Tables 6.3, 6.4, and 6.5 below list SoC percentages and peak scores for each individual at times 1, 2, and 3, respectively.

Table 6.3

Time 1: SoC Percentages and Peak Scores

	Stages of Concern							
Teacher	0	1	2	3	4	5	6	$\bar{x}$
A	20	49	40	54	31	37	49	40
B	23	86	57	77	66	57	43	58
C	54	46	17	20	37	26	23	32
D	46	57	20	91	66	46	51	54
E	34	100	80	46	66	94	31	64
F	11	9	14	31	23	29	11	18
G	49	37	3	54	86	26	9	38
H	29	26	43	69	69	34	23	42
I	49	46	23	69	54	37	37	45
J	9	37	51	54	37	34	40	37
K	17	77	60	69	77	54	17	53
L	9	20	3	23	26	17	6	15
M	51	77	57	80	66	69	31	62
N	31	74	29	69	66	51	29	50
O	11	23	11	51	26	29	11	23
$\bar{x}$ (N = 15)	30	51	34	57	53	43	27	42

Table 6.4

Time 2: SoC Percentages and Peak Scores

	Stages of Concern							
Teacher	0	1	2	3	4	5	6	$\bar{x}$
A	26	54	46	49	43	66	46	47
B	17	83	37	54	54	69	34	50
C	54	57	54	51	57	51	40	52
D	40	54	20	80	74	46	57	53
E	34	91	71	37	37	100	17	55
F	11	9	14	40	31	26	29	23
G	43	43	3	66	91	23	26	42
H	17	23	43	66	74	57	26	44
I	31	31	34	51	71	60	43	46
J	3	49	34	29	43	71	29	37
K	20	63	57	54	77	49	20	49
L	34	31	11	91	49	49	23	41
M	37	77	60	51	83	97	20	61
N	34	63	17	71	66	71	23	49
O	11	29	14	49	57	54	23	34
$\bar{x}$ (N = 15)	28	51	34	56	61	59	30	46

Table 6.5

Time 3: SoC Percentages and Peak Scores

	Stages of Concern							
Teacher	0	1	2	3	4	5	6	$\bar{x}$
A	14	49	31	34	31	69	34	37
B	11	80	23	34	31	86	23	41
C	69	69	94	100	71	69	66	77
D	31	49	20	71	77	46	60	51
E	34	83	60	29	20	100	6	47
F	11	9	14	40	40	20	37	24
G	23	26	3	43	86	57	34	39
H	23	23	37	63	77	66	29	45
I	29	29	37	43	71	74	43	47
J	3	54	31	17	43	80	14	35
K	26	54	54	46	74	51	29	48
L	40	14	11	66	63	60	43	42
M	29	63	60	31	83	100	49	59
N	37	51	6	57	60	97	20	47
O	20	31	14	40	69	69	29	39
$\bar{x}$ (N = 15)	27	46	33	48	60	70	34	45

Each teacher's peak score is ringed and it can be seen from Table 6.3 that 13 of the peaks at Time 1 relate to stages 0, 1, 2 and 3, reflecting most concern with self and (especially) management issues. By Time 2, that has reduced to 6 and, by time 3, to 3. By contrast, peaks at stages 4, 5, and 6, indicating greater concern about the effects of the CTP, increased from 4 at Time 1, through 11 at Time 2, to 14 at Time 3. As a group, then, and judging only by peak scores, the CTP teachers conform to the postulated developmental movement (from concerns about self to concerns about the impact of the innovation). Early concerns are reduced in intensity before later concerns emerge. The peak concerns for all 15 teachers are summarised in Table 6.6 below:

Table 6.6

Peak Concerns for 15 CTP Teachers

Stages of Concern:	0	1	2	3	4	5	6
Time 1	1	4	0	7	4	1	0
Time 2	0	2	0	4	6	5	0
Time 3	0	0	0	3	6	8	0

Note: These figures include double peaks for teachers whose scores were equally high at 2 stages

It is noticeable that not one respondent selected stage 6 (modifying the CTP) as the most intense concern at any period. In fact, it always remains a minor

concern. (Incidentally, this offers some corroboration of the Levels of Implementation finding [see Chapter 5] that the CTP was not generally perceived by teachers as an opportunity to modify according to each individual's sense of plausibility.) The only other stage which does not appear as a peak concern is Stage 2 (Personal), though as we shall see (in section 6.5.2), it is by no means negligible for some teachers.

The peak stage interpretation, showing that as a group teachers in this study conform with the hypothesised movement of concerns over time, is useful as a preliminary interpretation of the data. It does not, however, take account of the relative intensity of concerns at different periods. The following section examines each teacher's complete profile.

6.5.2 CTP Teachers' SoC Profiles

In order to attend to the relative intensity of teachers' concerns at times 1, 2 and 3, each teacher's complete profile is now considered separately. The interpretation is divided into three parts: firstly, the high and low stage scores are interpreted; secondly, a holistic summary of that interpretation is given; and thirdly, the teacher's responses to each stage are looked at to determine how well the items associated with different stages have been distinguished (i.e. a form of Q-sorting), and, therefore, how much confidence we are entitled to place in the SoCQ scores for each individual.

(An arbitrary rule of thumb for Q-sorting is desirable for purposes of consistency; at least 4 out of 5 responses for a particular Stage and Time must be scored at and below 3 (low), or at and above 4 (high), or from 2 to 5 (medium) for a good Q-sort; also, this must hold for at least 2 out of 3 Times). All of the interpretation is based on Tables 6.3, 6.4 and 6.5 above, on profile graphs (Figures 1 to 15 in Appendix 14), on the summary of each teacher's percentages by stage (Appendix 13), on the raw scores (Appendix 12), and on the raw scores arranged for Q-sorting (Appendix 15).

Teacher A

High and Low Stage Interpretation

At Time 1, teacher A has relatively intense concerns with regard to acquiring more substantive information about the CTP (Stage 1). The highest stage score is for management (Stage 3), but this is accompanied by a much higher than average score for Stage 6, indicating that although the teacher has concerns about logistics, time and organisation, he also has many ideas of his own about how to improve the CTP. At this Time, there is very little concern about the possible effects on students and relatively little about collaboration (Stages 4 and 5). Stage 0 is lower than average, but since all other stages have registered more perceptibly, it is possible to infer that teacher A has a fairly intense involvement with the program.

By Time 2, concerns at Stage 5 have taken a distinct upward curve and now form the highest peak by 12 percentage points. The second highest stage is Informational. Together, high scores at these stages suggest that Collaboration is better seen not as a desire to work with others or to coordinate teaching but as a strongly felt need to get ideas and find out more from others; this is lent some credence by the fact that Stage 6 concerns have diminished somewhat. Stage 0, 2 and 4 concerns have magnified at this Time, but Management concerns have abated.

By Time 3, the overall level of concern has dwindled. Only one concern is actually further aroused and that is at Stage 5. This is now 20 percentage points ahead of the next concern. However, since that second highest concern is still Stage 1, the interpretation of Time 2 holds, i.e. that Collaboration is best seen here as a need to gain ideas from others rather than working with them on an equal basis. Stage 6 concerns continue to fade, indicating (in conjunction with the high Stage 5 and 1 scores) that while teacher A began with ideas of his own which could potentially have modified the CTP according to his own lights, by Time 3, he was more dependent on the perceptions of others.

Holistic Appraisal

Overall, Teacher A seems to have yielded his own opinions to those of others, and appears to have required

the constant provision of information and ideas for teaching. Concerns in general showed an upward trend at Time 2, but waned again by Time 3. Thus the pattern of early concerns (0, 1, 2 and 3) decreasing while later concerns (4, 5 and 6) emerge did not materialise in this case.

Q-Sort

As can be seen in Appendix 16, looking from left to right, it is not immediately evident that responses to some stages are primarily low and primarily high for others. However, the Q-Sort rules of thumb indicate that Stage 0 is reasonably trustworthy (4 out of 5, 4 out of 5, and 5 out of 5 for Times 1, 2 and 3, respectively, are primarily low). Stages 2 and 5 are trustworthy at Times 2 and 3 (primarily high), and Stage 6 is trustworthy at Times 1 and 2 (also, primarily high). We may therefore place some confidence in the scores of Stages 0, 1, 5 and 6 and the interpretation expressed in the holistic appraisal. Stages 2, 3 and 4, however, are only trustworthy at one Time and must consequently be treated with more caution.

Teacher B

High and Low Stage Interpretation

At Time 1, Teacher B has low Stage 0 concerns but is overall markedly higher than average for the other Stages. As Hall, George and Rutherford (1977, p.53) point

out, this combination suggests intense involvement with the program. While all Stages from 1 to 6 are relatively high, Stage 1 (Informational) is considerably higher than the next highest (Management) which is in turn substantially higher than the other concerns.

At Time 2, overall concerns have declined; in fact all Stages subside except for Collaboration which jumps 12 percentage points. Although Stage 1 has lessened a little, it is still the highest concern by 14 percentage points. It is followed by Informational concerns. This combination begins to suggest that teacher B's Collaboration may, like teacher A's, be more a matter of securing ideas from others rather than contributing to a group. Stages 0, 2, 3, 4 and 6 have scaled down sharply.

By Time 3, it emerges quite clearly that the pattern that was forming at Time 2 has crystallised. Collaboration becomes the peak concern, closely followed by Informational concerns, corroborating the earlier, tentative interpretation that teacher B wishes to find out more about the CTP and that Collaboration for him is a means of fulfilling this wish. While Stage 1 and 5 concerns are high (80% and 86%), the other concerns have dimmed quite dramatically. The fact that Stage 2 is low while 1 is high indicates that the quest for more information is not accompanied by a sense of personal threat. The evaporating concerns about Renewal (Stage 6) suggest that teacher B originally had ideas about ways of improving

the CTP, but that these had faded.

Holistic Appraisal

Teacher B appears to have felt an increasing need to have information and ideas about teaching the CTP made available to him. He seems positively disposed towards the project (low Stage 2 with high Stage 1). Whatever opinions he may have had that might have modified his use of the CTP do not appear to have been sustained. Concerns in general showed a downward trend in all Stages except for Stage 1 (a slight downward trend, but remaining high) and Stage 5 (accelerating upwards from 57% through 69% to 86%). Thus, the postulated general pattern of development (a decrease in early Stages and an increase in later Stages) did not come about in the case of Teacher B.

Q-Sort

According to the Q-Sort rules of thumb, only Stage 2 is untrustworthy. Stages 0, 1, 3 and 6 are consistent at all three Times, and Stages 4 and 5 are consistent at two Times. Therefore, we may have a measure of confidence in the interpretation summarised in the holistic appraisal.

Teacher C

High and Low Stage Interpretation

At Time 1, Teacher C's peak concern is at Stage 0 and second highest concern at Stage 1. This suggests that he was very aware of the CTP and wanted to find out more about it. Coupled with the low Stage 2 concern, it may be

inferred that he was positively disposed to the project. Other concerns at this stage are minimal to moderate.

By Time 2, the picture has changed. Now all Stages are compressed to within a few percentage points of each other (except for a somewhat lower Stage 6). Although the peak concern is split between Stage 2 and Stage 4, it could have easily been any other stage. Everything has moved to a medium level of intensity.

By Time 3, again the picture has changed startlingly. Stages 0, 1, 4, 5, and 6 have been further kindled and are still within a few percentage points of each other. However, there has been a dramatic upsurge in Personal and Management concerns. Management concerns are at 100% closely followed by Personal concerns at 94%. There is a clear response pattern with teacher C: concerns are mostly low to start with; they all intensify at about the same rate, except for Stages 2 and 3 which acclerate ahead. Apart from noting that all concerns increase, what is salient for teacher C is his sense of inadequacy in relation to the CTP and how to deal with time and organisational matters.

Holistic Appraisal

Overall concerns increased from an average of 32 (lower than most teachers), through 52 (slightly higher than most) to 77 (18 percentage points higher than anyone). The only clearly focused concerns are Personal and Management issues. The postulated movement of

concerns (early concerns lowering, later concerns increasing) did not materialise in this case.

Q-Sort

Q-Sorting indicates that responses to Stages 0 and 1 were relatively untrustworthy and that those to Stages 2, 3, 4, 5 and 6 were relatively trustworthy (at all 3 Times). The response pattern is clear and suggests that teacher C attended to the questionnaire very carefully and distinguished well between Stages. We may have increased confidence, therefore, that the interpretation given in the holistic appraisal is based on reasonably consistent data.

Teacher D

High and Low Stage Interpretation

At Time 1, teacher D has substantial concerns at all Stages except Personal, which suggests that he is very interested in the project and is unlikely to be put off by feelings of personal inadequacy to carry out what the CTP demands. By far the most intense concern is Management. The second highest concern is with regard to the likely impact on pupils. A reasonably high Stage 6 percentage indicates that teacher D may have ideas which are independent of the CTP, and which may later lead to some form of modification.

At Time 2, Stage 2 remains at the same minimal level (20%), but all of the other early concerns decrease. Even

with this decrease, however, Management is still the peak concern. The late Stage concerns all increase, except for Collaboration which remains static. The second highest concern is still Consequences for pupils.

By Time 3, Consequences has continued its upward trend and is now the peak concern. The second highest Stage is Management, but this is falling substantially. Again, Stages 0, 1, and 3 decline, while 2 remains minimal; by contrast, Stages 4 and 6 climb steadily, 5 remaining static. Concerns appear to be shifting towards the effects of the CTP and ideas for improving it.

Holistic Appraisal

Overall, teacher D conforms to the postulated developmnental movement of concerns: early concerns recede as later concerns emerge. Total averages across Stages at all 3 Times are only slightly above the mean for the group of 15 CTP teachers. Although teacher D has marked, though diminishing Management concerns, he also has steadily rising Stage 6 concerns, suggesting that as his own ideas gain a foothold, problems pertaining to organisation can be overcome. The most striking developmental movement is in relation to Consequences, however.

Q-Sort

Only Stages 2 and 3 are trustworthy according to the Q-Sort rules of thumb. In fact, of the 21 cells in a Stage by Time matrix, only 6 are trustworthy. The main

reason for this appears to be that teacher D has an extreme response tendency: 79% of his responses are either 0, 1 or 7. He either has a concern or he has not; there are few shades between. Given that the rules of thumb allow some latitude, respondents who use a greater range of scores are more likely to appear trustworthy. It is worth retaining the same rules, however, because a respondent who gives a generally high rating to a set of questions which are supposed to tap the same concern is to be preferred to a respondent who apparently fails to recognise (implicitly) the similarity of the questions and therefore careers between high and low extremes. We have little choice but to treat the interpretation of teacher D's SoCQ responses with due caution.

Teacher E

High and Low Stage Interpretation

At Time 1, teacher E's mean percentage across Stages is the highest of all the CTP teachers. The most intense concern is in regard to Information (100%), which coupled with a second highest concern with Collaboration (94%) indicates that the teacher is extremely interested in the program and wants to find out more about it and to know what other CTP teachers are doing in the classroom. The low score for Stage 6 suggests that teacher E has few ideas of his own which might later conflict with or modify others' perceptions of the project. Personal

concerns are 20 percentage points higher than anyone else's at this Time; this would contribute to an interpretation that Collaboration is best seen as a means of obtaining ideas from others rather than contributing.

At Time 2, Stage 6 is fading fast (only 17%), while Stage 1 decreases slightly (91%) and is the second highest concern, and Stage 5 increases to 100%. Thus, the picture of teacher E as interested in the CTP but dependent on the perceptions of others is increasingly clear. All concerns wane at this Time, except Collaboration (and Awareness does not move), but the most dramatic decline (of 29 percentage points) is Consequence. Recalling that Time 2 relates to "after a few weeks of using the CTP", it may be that the steep decline in concern about the impact of the CTP on pupils is a reflection of the overriding concern with gaining information and ideas with regard to the project.

By Time 3, the trends observed at Time 2 continue; Collaboration is still the highest concern at 100%, followed by Informational concerns (down to 83%), and Refocusing (Stage 6) is virtually fallow (6%). Consequence continues its sharp decline (to 20%). Personal concerns have become slightly less salient but are still marked (60%).

Holistic Appraisal

Teacher E's concerns do not move in the directions anticipated by concerns theory. At Time 3, (toward the

end of the teacher's association with the CTP), concerns about the effects of the project on pupils and ideas for modification, far from proliferating, have almost disappeared. The clear, overriding concern is to find out more about the CTP and to get ideas about how to implement it.

Q-Sort

All Stages are trustworthy, except for 0 and 3, which are not central to the interpretations offered above. This permits increased confidence that our interpretation is based on reasonably consistent data.

Teacher F

High and Low Stage Interpretation

At Time 1, teacher F's overall concerns are the second lowest of the 15 CTP teachers (18%). When first aware of the CTP, it is clear that concerns are minimal, but Management is the peak concern, followed by Collaboration. It would appear that although concerns in general are subdued, matters pertaining to time, energy and organisational issues are uppermost, along with an interest in how other CTP teachers are implementing the project.

At Time 2, the highest concern is still with Management matters, but the second highest concern is with Consequence, as Collaboration declines in intensity. Stage 6 concerns have jumped from 11% to 29%. Stages 0,

1, and 2 remain unchanged. Overall concerns are at 23%, which is 11 percentage points lower than anyone else at this Time.

By Time 3, Management concerns are still uppermost but are now equalled by Consequence concerns. The only other increase is Refocusing. Collaboration concerns continue their decline and Awareness, Information and Personal remain minimal and unchanged. It appears that teacher F did not take an active interest in the CTP at any Stage and made little effort to find out more about it. The tailing up of Stage 6 concerns indicate that he may have had ideas of his own which increasingly come to compete with CTP perceptions. If this interpretation is correct, the peak Management concerns and the steady increase in Consequence concerns may be due to a less than positive attitude towards the project and a view that pupils would be better served by another approach.

Holistic Appraisal

Concerns at all 3 Times are generally low. The fact that they are low and change very little suggests a lack of interest in the CTP and the tailing up of Stage 6 concerns indicates that other perceptions increasingly come to the fore.

Q-Sort

Stages 0, 1, 2 and 5 are relatively trustworthy, but Stages 3, 4 and 6 are not. This seems to strengthen the lack-of-interest interpretation in that we cannot even

have moderate confidence in the apparent increase in Consequence concerns. The interpretation that teacher F's own ideas increasingly came to the fore must be treated with caution, and the apparently peak Management concerns cannot be interpreted with confidence.

Teacher G

High and Low Peak Interpretation

At Time 1, the clear peak concern is Consequence (by 32% over other concerns). This is accompanied by moderately high Stage 0, 1 and 3 concerns. Stage 2 concerns are virtually non-existent (3%) and Stage 6 concerns barely perceptible (9%). This configuration suggests an interest in the CTP, an attitude unclouded by concerns about personal adequacy, few notions that the approach might be modified, and an overriding interest in the potential impact on pupils' learning.

At Time 2, the concern about Consequence is still much higher than anything else. Stage 0, 1 and 2 remain fairly stable, though Management concerns increase and Stage 6 concerns now register clearly. It seems that with a few weeks experience of the CTP, the main changes are that teacher G begins to develop ideas that might come to modify the approach, and that initially, there are more problems with Management than anticipated at Time 1.

By Time 3, Stage 0 and 1 concerns have dropped considerably as have Management concerns. Stage 2 remains

at 3%, but Stage 6 has jumped to 34%. Stage 5 becomes the second highest concern and Consequence remains clearly the peak area of concerns. Early concerns, therefore, if they were not already very low, have all faded. By contrast, late concerns have either remained high or increased markedly. Thus, the postulated developmental movement of concerns is realised in the case of teacher G.

Holistic Appraisal

The movement of concerns follows the hypothesised pattern. After sorting out difficulties relating to Management, and acquiring information about the CTP, teacher G increasingly develops his own ideas, wishes to collaborate with colleagues, and retains a principal concern with the impact of the project on pupils.

Q-Sort

All Stages are trustworthy except for 0, 1 and 3 which means that we cannot be confident that these Stages actually decreased in intensity.

Teacher H

High and Low Peak Interpretation

At Time 1, there are low to moderate concerns at all Stages except for Management and Consequence (at 69% each). When teacher H is first aware of the CTP, it seems that there is general interest accompanied by particular concerns about time, energy and organisational matters and also about the potential effects of the approach on

pupils.

At Time 2, the early Stages recede (Stage 2 remains the same), and the late Stages all assume greater prominence. The peak concern is Consequence, and Management, though declining, is still the second highest concern. The most marked increase is with Collaboration which, given the downward trend of early concerns and the high Consequence concern, can be interpreted as a wish to work with others to ensure beneficial effects on learners.

By Time 3, all of the late concerns have continued to rise, while early concerns either remain low or subside even further. Consequence is still the peak concern, but Collaboration has by now taken over from Management as the second highest concern. Stage 6 concerns have continued to rise (though only slightly). The clearest movements over the 3 Times are the emergence of Collaboration and the dominance of Consequence concerns.

Holistic Appraisal

Teacher H follows the postulated movement of concerns over time. The lowering of early concerns makes way for the increase in late concerns. The strong interest in the impact of the CTP is less and less burdened by practical Management concerns; it may be that solutions to Management problems were found through working with others (Collaboration) and through the slow

emergence of the teacher's own ideas for modifying the approach (rising Stage 6).

Q-Sort

The only Stage which appears to be untrustworthy is Stage 2. Since the scores in this Stage were moderate and contributed little to the overall interpretation, some confidence in the Holistic Appraisal is possible.

Teacher I

High and Low Peak Interpretation

At Time 1, teacher I is principally concerned about Management issues and about the potential impact of the CTP on pupils. Personal concerns are the lowest of all, suggesting that the interest in the project implied by moderate Stage 1 concerns will not be hampered by a sense of uneasiness about personal adequacy in relation to the approach.

At Time 2, all Stages change in the postulated directions, i.e. early concerns curve downwards, while late concerns rise. The only exception is Stage 2, which increases from 23% to 34%, which is an anomaly that is difficult to account for. The peak concern is now for Consequences and the second highest concern is for Collaboration. Stage 6 tails up slightly, indicating that teacher I's own ideas in relation to the CTP are present after short experience of the project.

By Time 3, the same trends observed at Time 2

continue. Concerns either remain as they were or move in the hypothesised directions. The peak concern is now Collaboration, though closely followed by Consequence issues. The tailing up of Stage 6 noted at Time 2 has levelled out, intimating that teacher I's ideas are unlikely to lead to revision of the CTP approach. Stage 2 concerns are increasingly aroused; increasing Personal concerns are difficult to square with declining Stage 1 and 3 and increasing Stage 4 concerns.

Holistic Appraisal

Apart from the anomalous Stage 2 movement, early concerns tend to decrease as late concerns escalate. Teacher I is most concerned about the effects of the CTP on pupils and about working in conjunction with colleagues. It appears unlikely that his own ideas will become very prominent.

Q-Sort

The anomalous Stage 2 score cannot be explained away as an untrustworthy set of responses to the SoCQ. Stages 2, 3 and 4 are the only trustworthy areas, while 0, 1, 5 and 6 provoke more random responses. Too much is uncertain for even moderate confidence in the interpretation offered in the Holistic Appraisal. About all that can be said is that as Managment concerns decline, Consequence concerns emerge. Beyond that, inference is precarious.

Teacher J

High and Low Peak Interpretation

At Time 1, the peak concern is Management, followed by Personal. All of the other concerns are medium except for Awareness which is very low (9%). Hall, George and Rutherford (1977, p.53) state that a low Stage 0 score with other Stages being substantial testifies to an intense concern about the project.

At Time 2, the already minimal Stage 0 concern almost vanishes (3%). Management concerns, previously the most intense, have subsided markedly (from 54% to 29%). Stage 5, in the meantime has jumped from 34% to 71%, making it the peak concern. This is now followed by Stage 1, which has risen from 37% to 49%. The combination of high Stage 1 and 5 scores signifies a need to gather information such that Collaboration is probably better seen as a means of fulfilling this need rather than as a wish to contribute. Stage 6 concerns decline as do Personal anxieties.

By Time 3, Stage 5 is still the peak concern and is still followed by Stage 1, (80% and 54%, respectively). Stage 6 has dwindled to 14%, and Management to 17%. Personal concerns continue to wane and Consequence concerns remain moderate. By now, there are only 3 Stages that are important: Stages 1, 4 and 5. The desire to acquire information is paramount, but is accompanied by a concern about the effects of the project on pupils'

learning.

Holistic Appraisal

The principal concern is to find out more about the CTP. Concern is also high that pupils should benefit. Teacher J's own ideas in relation to the project decay over time. The hypothesised movement of concerns is not adhered to in this case.

Q-Sort

Stages 0, 2, 4, 5, and 6 are trustworthy, but Stages 1 and 3 are responded to inconsistently. Therefore, it is possible that teacher J does not have high concerns about Information. This in turn means that Stage 5 could be a genuine wish to contribute with colleagues. This part of the Holistic Appraisal has to be treated with caution.

Teacher K

High and Low Peak Interpretation

At Time 1, the lowest concerns are with Awareness and Renewal, while all other Stages are high. This argues that teacher K is intensely involved with the project but brings few ideas of his own to it. The peak concerns surround Information and Consequence.

At Time 2, the only Stages to increase are Awareness and Refocusing, but they remain peripheral. The central concerns are still with 4 (77%) and 1 (63%). Stage 4 has not changed, but 1, 2, 3, and 5 have all abated, but only

slightly, and they remain substantial.

By Time 3, the pattern initiated at Time 2 has become established. Again, Stage 0 and 6 concerns increase, but so slightly that they are still minimal. Stage 1, 2, 3, and 4 concerns decrease, but so slightly that they remain substantial. The peak concern is still with regard to Consequence; this is followed by Information and Personal concerns. Stage 5 actually increases, but only by 2 percent. The only clear focus is the concern about Consequence. Apart from this, and the minimal Stage 0 and 6 concerns, it appears that teacher K has multiple concerns, and there is no indication that some will subside and others emerge.

Holistic Appraisal

For teacher K, there appears to be no clear developmental movement of concerns. A general, unfocused set of substantial concerns is only broken by a sustained high peak in relation to Consequence and a minimal concern about Stages 0 and 6. Interpretation beyond this would not be justifiable.

Q-Sort

Stages 0, 2, 3, 4, and 6 are all relatively trustworthy. Only Stages 1 and 5 are particularly suspect. This Q-sort does not materially alter the minimal interpretation offered in the Holistic Appraisal.

Teacher L

High and Low Peak Interpretation

Concerns are extremely low overall, ranging from 3% to 26%, with a mean of 15%. This is the lowest of all 15 CTP teachers. The peak concern is with Consequence, followed by Management. At this time, it appears that if teacher L thought about the CTP atall, it was only to wonder about organisation and potential effects on learners.

At Time 2, the picture changes dramatically. The mean level of concerns is now up to 41%, only a few percent lower than the average for the 15 teachers. Every concern increases in these early days of using the CTP. The slightest increase is for Personal concerns, up to a mere 11%, indicating that the teacher is not atall hindered by worries relating to personal adequacies. Concerns about Consequence accumulate rapidly (from 26% to 49%) as do Collaboration concerns (from 17% to 49%). Stage 6 tails up considerably (from 6% to 23%), suggesting that teacher L's own ideas may become more prominent. The most striking upward trend, however, is with Management (from 23% to 91%). This is the peak concern now, followed by Stages 4 and 5. Management concerns are 42% higher than anything else and are clearly salient for teacher L.

By Time 3, early concerns tend to decline or remain minimal. Most dramatically, Management concerns have subsided from 91% to 66%, and though still the peak

concern, the trend is clearly downwards. Late concerns are all clearly rising. Consequence is now the second highest (63%). Stage 6 has jumped again, and is now 43%, from which it might be inferred that the teacher's own ideas are likely to make a perceptible impact on her implementation of the CTP.

Holistic Appraisal

Teacher L evinces very little interest in the CTP when she is first aware of it, but once she commences her association with the project teaching, her interest increases dramatically, and the most obvious concern is to do with organisational matters. Towards the end of her association with the CTP, she appears to have obtained all the information she needs, Management concerns recede sharply, and concerns about Consequence, Collaboration and Refocusing are increasingly roused. She appears to have strong ideas of her own for implementation.

Q-Sort

Stages 1, 2, 3, 4, and 6 are relatively trustworthy. Only Stages 0 and 5 are responded to inconsistently and these have little effect on the interpretation offered in the Holistic Appraisal.

Teacher M

High and Low Peak Interpretation

At Time 1, teacher M is most concerned about Management, followed by Information. All Stages, however,

are fairly high and although this makes comparisons of relative intensity difficult, it does suggest that the teacher is intensely concerned about the CTP.

At Time 2, and with a little expereience of actually teaching on the project, teacher M's concerns shift somewhat in the hypothesised direction. While both the peak and second highest concerns at Time 1 were early concerns, by Time 2, they are late concerns, Collaboration first and Consequence second. Management concerns diminish (from 80% to 51%). Stage 6 is rather more subdued. Stage 1 has levelled out, and Stage 0 has lessened. Stage 2 has increased but only very slightly.

By Time 3, the hypothesised movement of concerns seems to be regular. Stages 0, 1, and 3 recede, and 2 has levelled out. Stage 4 levels out at 83% and is the second highest concern, while Stage 5 becomes the peak concern (100%) Stage 6 leaps from 20% to 49%, indicating that teacher M's own ideas are likely to have a bearing on her implementation of the CTP. Although Stages 1 and 2 have declined or levelled out, they are still high (63% and 60%, respectively). This intimates that she is concerned about her adequacy to teach satisfactorily on the project and that her high score for Collaboration may be at least partly influenced by a wish to find out more about the CTP than a desire to contribute.

Holistic Appraisal

Teacher M's early concerns do recede but the need to

acquire further information always remains high as do concerns about personal adequacy. Late concerns come to assume greater prominence, though the desire for Collaboration may be influenced by a perceived need to get ideas for teaching. The emerging concern about Stage 6 may imply that teacher M's own ideas have an effect on her implementation. Her concerns overall are well above average at all 3 Times, possibly indicating that she has not resolved certain difficulties in relation to the CTP.

Q-Sort

All Stages are fairly trustworthy except for Stage 0. There is no reason, therefore, to modify the interpretation given in the Holistic Appraisal.

Teacher N

High and Low Peak Interpretation

At Time 1, concerns about Information, Management and Consequence are fairly high, Collaboration is moderate and Awareness, Personal and Refocusing concerns are relatively low.

At Time 2, Management and Collaboration concerns are highest, followed by Consequence and Information. Other Stages are relatively low.

By Time 3, Collaboration concerns are strikingly clear (97%), followed by moderate concerns about Consequence, Management and Information. Since Stage 1 remains quite high, it is possible that the high Stage 5

score is partly coloured by a wish to obtain ideas from others. Personal concerns have by now practically disappeared and Refocusing concerns have faded.

Holistic Appraisal

Apart from the very low Personal concerns, the negligible Refocusing concerns and the very high Collaboration concerns, no clearly focused concerns are discernible over time for teacher N. The focus on Collaboration may due in part to a wish to get ideas for CTP teaching from others. Teacher N seems unlikely to bring many of his own ideas to the project.

Q-Sort

Stages 0 and 4 are relatively untrustworthy; all other Stages seem quite consistent. Thus, the interpretation presented in the Holistic Appraisal need not be modified.

Teacher O

High and Low Peak Interpretation

At Time 1, all responses evince very low concerns except for Stage 3 (51%). At this Time, it appears that the teacher has relatively interest in the CTP, but anticipates organisational problems.

At Time 2, Stages 0, 1, 2 and 6 remain low though apart from Stage 0, they increase slightly. Stage 3 declines slightly, and Consequence becomes the peak

concern, having jumped from 26% to 57%. The second most important concern is Collaboration. Although Stage 6 remains low, it has markedly increased from 11% to 23%.

By Time 3, Stages 0, 1 and 2 are still low, Stage 3 has reduced to 40% and Stages 4 and 5 have become quite intense concerns (both 69%). Stage 6, although still fairly low, has continued to rise and might be considered likely to influence teacher O's implementation of the CTP.

Holistic Appraisal

The movement of Stages 0, 1 and 2 are not substantial, so we might infer that teacher O soon felt she had acquired all the information she required, and that she was untroubled by concerns about personal adequacy. Management concerns remain substantial, though they do recede perceptibly. The most marked changes over time relate to concerns about the effects of the CTP on learners and a wish to work with others on the project. The steady tailing up on Stage 6 suggests that teacher O may have ideas of her own which might influence her implementation of the CTP.

Q-Sort

All Stages are relatively trustworthy. In fact, out of the 21 categories in a Stages by Time matrix, 20 are responded to consistently. There would appear to be no reason to modify the the interpretation offered in the Holistic Appraisal.

Note: The Stage by Time matrices used for the Q-sorts are presented in Appendix 16.

6.5.3 Interactions between SoC and RT/NRT

In chapter 5, a fairly strong association between Level of Implementation and whether a teacher was regular (RT) or non-regular (NRT) was reported. Since this has important implications for the external validity of the CTP, the RT/NRT issue was thought to be worth examining from the perspective of Stages of Concern. To this end, an analysis of variance was carried out.

Although it might be argued that a one-way analysis would be appropriate, this would only inform us about the relationship between the actual percentage means and RT or NRT status. More precision would be possible if stages were taken into account as well as percentages so as to augment reflection about differences between the 7 stages as they relate to the RT/NRT distinction. Also, given that an analysis of variance, in Guilford and Fruchter's words, provides us with "only an overall answer regarding the significance of a whole collection of differences between means" (1978, p.235), it would be unnecessarily restrictive to preclude the consideration of specific differences at the 7 stages. Potentially more could be gained by applying a multifactor analysis of variance with the percentages as the response variable and the two

sources of variation being the 7 Stages of Concern and the RT/NRT distinction.

But before the analysis, the cautions suggested by Woods, Fletcher and Hughes (1986) are worth heeding. They note that in linguistic research a common problem is caused by data in the form of proportions or percentages in that they could well fail to be normally distributed (a principal assumption in ANOVA studies), especially if some of the scores fall outside the 10% - 90% range (as is the case with our data). Woods, Fletcher and Hughes (1986, p.220) recommend that in such circumstances the scores be re-scaled so that they "will be normally distributed and have constant variance", and that the requisite tool is arcsine transformation. The duly transformed scores are presented in Appendix 17.

Following the arcsine transformation, a multifactor analysis of variance was run on the SoC data with the results shown in Tables 6.7, 6.8 and 6.9.

Table 6.7

Multifactor Analysis of Variance for Time 1

Source of Variation	Sum of Squares	df	Mean Square	F-ratio	Sig. level
MAIN EFFECTS	6508.0747	7	929.7250	5.368	.0000
SoC	5451.8667	6	908.6444	5.246	.0001
RTNRT	1056.2080	1	1056.2080	6.098	.0154
2-FACTOR INTERACTIONS	1307.3541	6	217.89235	1.258	.2847
RTNRT SoC	1307.3541	6	217.89235	1.258	.2847
RESIDUAL	15760.705	91	173.19456		
TOTAL (CORR.)	23576.133	104			

Table 6.8

Multifactor Analysis of Variance for Time Two

Source of Variation	Sum of Squares	df	Mean Square	F-ratio	Sig. level
MAIN EFFECTS	8833.9621	7	1261.9946	10.798	.0000
SoC	7590.3238	6	1265.0540	10.824	.0000
RTNRT	1243.6383	1	1243.6383	10.640	.0016
2-FACTOR INTERACTIONS	2378.1632	6	396.36053	3.391	.0046
SOC RTNRT	2378.1632	6	396.36053	3.391	.0046
RESIDUAL	10635.932	91	116.87837		
TOTAL (CORR.)	21848.057	104			

Table 6.9

Multifactor Analysis of Variance for Time Three

Source of Variation	Sum of Squares	df	Mean Square	F-ratio	Sig. level
MAIN EFFECTS	11426.826	7	1632.4037	9.612	.0000
SoC	9329.257	6	1554.8762	9.156	.0000
RTNRT	2097.569	1	2097.5690	12.352	.0007
2-FACTOR INTERACTIONS	2184.6779	6	364.11299	2.144	.0558
SoC RTNRT	2184.6779	6	364.11299	2.144	.0558
RESIDUAL	15453.886	91	169.82293		
TOTAL (CORR.)	29065.390	104			

It may be seen that at all three times the main effects of Stages of Concern and RT/NRT are significant at least at the .05 level of significance. Thus it would seem that differences existed between the 7 stages and also that overall there were real differences between RTs and NRTs. It may be seen, in addition, from Tables 6.7, 6.8 and 6.9 that although no significant interaction exists between SoC and RT/NRT at Time 1, at Time 2 the interaction is significant, and it only just misses being significant at Time 3. Practically, it would seem wise not to overlook the specific differences between RTs and NRTs at different Stages.

Since all of the possible comparisons are of interest, t-tests cannot be used with confidence because the comparisons are not independent of each other. And although, as Woods, Fletcher and Hughes (1986, p.210) point out, there are theoretically correct procedures, they are very complex. Given that our data are likely to contain error (reliability and validity have not been satisfactorily established), it would be inappropriate to use such refined analytical tools (see Davies 1984, p.112). Woods, Fletcher and Hughes recommend instead that a rule of thumb be adopted: "provided that the ANOVA has indicated a significant difference between a set of means, calculate the standard error s* for the comparison of any pair of means by:

$$s^* = \sqrt{\frac{2 \times \text{residual mean square}}{n}}$$

where n is the number of observations which have been averaged when calculating each mean. Then find the difference between each pair. If the difference between a pair of means is greater than 2s*, take this as suggesting that the corresponding population means may be different. If the difference in two sample means is greater than 3s*, take this as reasonably convincing evidence of a real difference" (1986, p.210).

Calculating the differences in pairs of means in this way, the results presented in Table 6.10 were obtained.

Table 6.10

Differences between RTs and NRTs at 7 Stages and 3 Times

	TIME 1		TIME 2		TIME 3	
Stage	diff	size	diff	size	diff	size
0	9.45	>3s*	9.91	>3s*	8.75	>3s*
1	17.59	>3s*	14.59	>3s*	17.11	>3s*
2	13.11	>3s*	20.04	>3s*	25.73	>3s*
3	6.00	>3s*	7.13	>3s*	4.98	>2s*
4	2.54	>s*	4.04	>2s*	6.72	>3s*
5	12.43	>3s*	21.32	>3s*	18.81	>3s*
6	6.16	>3s*	0.20	<s*	2.09	>s*

At Time 1, s* = 1.81; at Time 2, 1.49; at Time 3, 1.79.

Where the size of the difference is underlined, this indicates that the higher mean was obtained by the NRTs; in all other cases, the RTs obtained the higher means.

It may be seen at a glance that nearly all of the differences are probably real (according to the rule of thumb of Woods, Fletcher and Hughes). Also, in the majority of cases, the higher means are obtained by the RTs. This would suggest that they were generally more concerned overall and that their CTP experience was more markedly characterised by anxiety than was the NRTs'.

It is noticeable that the earlier concerns (with self) are far more substantial for the RTs than for NRTs, especially at Stage 2, over all three times. At Stage 3 (management), the NRTs are initially more concerned, but by Time 3 they have been overtaken by the RTs. NRTs are more concerned about the effects of the CTP on pupils than are the RTs and the difference increases as time goes on. The substantially higher concerns felt by the RTs at Stage 5 (Collaboration) can be interpreted as a need to obtain ideas from others rather than contribute any, given their high concerns with simply coping (i.e. the high early concerns). At Stage 6, no clear differences emerge over time, and this probably reflects in the RTs a sense of not being in a position to think beyond what they can hardly manage in the first place; in the NRTs, in view of the Levels of Implementation findings of chapter 5, it would appear that a wish to conform to perceived ideas of CTP practice outweighed concerns about a shift of focus. All of these results and interpretations are consistent with the differences

between RTs and NRTs reported in chapter 5.

However, to end this section on a note of caution, it would be sensible to consider the ANOVA results tentatively. This is primarily because since (as already mentioned) the reliability and validity of the data-gathering instrument have not been satisfactorily established and since self-report is notoriously unreliable, the data themselves may include a good deal of error, and thus whatever statistical analysis is brought to bear, we must constantly bear in mind the nature of the data themselves and regard the ANOVA results as a set of indications rather than confirmations. (On the other hand, even a casual glance at the differences in means indicates that marked differences between RTs and NRTs probably do exist at the different Stages).

6.6 Summary and Discussion

The motivation for the data collection and analysis of the present chapter has been to facilitate extrapolation from the CTP to other circumstances (the notion of external validity advanced by Cronbach 1982, among others). The focus of this chapter has been on the concerns felt by teachers when they took part in the innovatory Bangalore project. By attending to the concerns of individual teachers, it is possible for projects similar to the CTP to be set up with greater

expectations of the kinds of difficulties individuals are likely to face. (The more similar the nature of the putative project, or the more similar the setting and socioeconomic factors, the types of teachers and students, and so on, the less distance the extrapolation has to travel, and the greater the likelihood of its being correct).

For our purposes, the Stages of Concern Questionnaire developed by Hall, George and Rutherford (1977) was selected as it addressed the issues of interest and, equally importantly, it had a track-record of reliability and validity. Certain changes were made, most of them minor, but others (viz. the use as a retrospective rather than introspective measure) more substantial. The consequent effects on reliability and validity are unquantifiable, but it was argued that the adoption and adaptation of an established instrument was preferable to developing one that could not be validated atall (for practical reasons). Inevitably, however, the results, the rational interpretations and the findings suggested by statistical analysis must be treated with caution.

Although the main thrust of the chapter has been towards the meticulous description of individuals (section 6.5.2), summaries and interpretations of the data are offered by the focus on peak-stage scores (section 6.5.1) and on the interactions between the 7 Stages of Concern and RT/NRT distinctions (section

6.5.3).

General findings from the peak stage score interpretation were (i) that there was an overall movement of concerns from self at Time 1 to impact at Time 3, (ii) that early concerns tend to dwindle before later concerns emerge, and (iii) that Stage 6 was always a minor concern.

With respect to the investigation into the relationships between the 7 Stages of Concern and RT/NRT distinctions, it was found that (i) real differences probably exist between RTs and NRTs, (ii) RTs are more concerned overall than NRTs, (iii) early concerns are far more marked (especially Stage 2) for RTs than for NRTs, (iv) NRTs are more concerned about effects on pupils than are RTs, (v) 'Collaboration' for RTs tends to mean the need to gather ideas from others rather than contribute their own, and (vi) Stage 6 (refocusing), which is low overall, is probably best interpreted as simply beyond the RTs (who were more concerned with coping), and perhaps undesirable for the NRTs (who, as the findings of chapter 5 indicate, were inclined to adhere closely to CTP perceptions).

Perhaps the single most pertinent finding for the purposes of external validity of the CTP evaluation is that with regard to the distinctions between RTs and NRTs, all of the indications emerging from the Stages of Concern inquiry are congruent with those of the Levels of

Implementation study. Thus, in spite of all the necessary caveats that characterise both investigations, such findings as there are seem to triangulate.

CHAPTER 7

CTP TEACHERS' TREATMENT OF PUPIL ERROR

7 CTP Teachers' Treatment of Pupil Error

This chapter considers CTP teachers' treatment of pupil error. From a data-base of 21 CTP lesson transcripts, distinctions between linguistic and content error are investigated and a number of sub-divisions are explored. The actual treatments are compared with the CTP statements about the ways that error should be handled to accord with stated perceptions about the differences between 'planned' and 'natural' language control.

7.1 Rationale

Prabhu (1982, p.5) makes a distinction between 'planned' and 'natural' language control by teachers. Essentially, planned control refers to a focus on form and natural control to a focus on meaning. Planned control implies a prior decision about what language is to be used, while natural control relates to an ongoing judgment by the teacher of what the learner is able to manage. The central tenet of the CTP is that grammar is best learnt without a focus on form but with a focus on meaning, so the distinction between planned and natural control is a crucial one in trying to understand differences between CTP teaching and other, specifically structure-based, forms of teaching. A major stumbling block has been that planned and natural control may manifest themselves in the same tokens, and that since

actual utterances may be the same, the distinction lies in the intention of the speaker. An example (if rather a long one) is in order.

At a seminar held in the RIE in Bangalore in 1980, a transcript of a CTP lesson was presented to seminar participants. This transcript is displayed in Appendix 18. The lesson extract deals with distances on a map and the teacher asks the pupils how far one town is from another. Almost the only question the teacher asks is "How far is [X] from [Y]?", and he proceeds to swap roles so that the pupils ask him about distances. In all, the teacher asks the question "How far...?" 15 times and the pupils ask it 7 times. This is followed by an exercise in which pupils must answer 5 questions of the form "How far...?", and then must write 5 questions for which the answer (the distance) is given.

(In the discussion that follows, all of the quotes are from RIE Bulletin 4 (i), 1980, pp.48-53).

After considering the transcript, the very first comment by a participant is: "One thing is very clear. One structure is drilled. Is that the intention?" He follows this question with another: "In the second part of the task, was there not an attempt to get back the structure from the students?" Prabhu responds that that "was never the intention". Another participant claims that if he were doing a structural lesson, he would do it in the same way. Keith Johnson observes: "If the teacher has predetermined what structure he is going to deal

with, then this is a structural lesson. And it seems to me that this is the case. The teacher selected the map with 'how far...' in mind", to which Prabhu retorts, "It came in naturally".

A little later, Johnson takes the view that if it is not a structural lesson it could just as easily be a notional lesson about 'asking about distances' and adds: "The answer in my view depends upon what it is that the teacher is trying to do when he goes into the classroom, whether to practice 'how far...' and other ways of asking about distances, or to give an interesting communicative activity". Prabhu rejoins that the intention to teach form would be noticeable: "The lesson would reflect it".

The problem is that for the seminar participants the CTP lesson reflects structural or possibly notional teaching. The debate apparently ends with all parties entrenched in their views.

It seems probable that Johnson is correct when he states that the differences are in the minds of the teachers and not in the tokens of their classroom talk. But there seems no tangible way of getting at intentionality; and if intentionality is the only difference, does that make any difference to the learners?

Proponents of methods have often been particularly difficult to pin down; few are prepared to state precisely what behaviours and language will occur that

are distinct. The very vagueness of methods may well have contributed to their survival. Fortunately, however, in relation to the CTP, Prabhu has made some fairly clear statements about what is acceptable and what is not with regard to teacher treatment of pupil error, and since this is an important part of classroom language control, it is worth examining.

7.2 CTP Attitudes to the Treatment of Error

Prabhu (1982, p.5-6) describes 4 elements of the kind of 'incidental' correction that he believes is appropriate to a meaning-focused classroom:

> (a) Incorrect language from learners is corrected (i.e. rephrased, restated, or drawn attention to) in roughly the same way that interested adults do with children - or the subject-teacher in an English-medium class does in teaching his subject.
>
> (b) This is done more in the context of writing (either on the blackboard, as part of the pre-task, or on paper in performing the task) than in oral work, as being more natural in that context.
>
> (c) All such attention to language is limited to facts (as against generalisations) and treated as contributory to the successful performance of the task on hand.
>
> (d) Learners' work is always marked for content, not correctness of language, though errors of language are corrected (as far as they can be, in the time available). Learners are not asked to rewrite in the light of the observations made.

The reference to correction as "roughly the same" as that of interested adults to children is not perhaps very helpful since interested adults vary considerably , and

Prabhu is only ready to commit himself to an approximation anyway. The reference to the concentration of error treatment in written work would have been relevant but recordings of lessons that were available to the present inquiry do not permit observation of corrections made in students' notebooks, and as most of the recordings are audio, it is not even possible to know what was written on the board. We are thus restricted to oral treatment of error. Thus, the only reasonably tangible description is that error treatment is limited to 'facts', that 'generalisations' are unacceptable, and that the treatment is necessary for the completion of the task.

In Prabhu (1987), there are further statements about the treatment of error:

> The teachers made the correction on the black-board, or told the learner who was writing what to change, but did not attempt to follow up an error with an explanation or other examples of the same kind (p.62).

This elaborates somewhat the kind of generalisation that is unacceptable: no explanation and no exemplification.

Prabhu (1987) proceeds at some length:

> It seems useful to useful to call such language-repair 'incidental correction' and to distinguish it from 'systematic correction', which involves a <u>larger</u> interruption of ongoing activity to focus learners' attention on an error that has taken place by providing an explanation or a set of other such instances in the hope of preventing a recurrence of the type of error it represents. Systematic correction also involves making the errors noticed in one lesson the basis of some planned work in the classroom in a subsequent lesson or anticipating particular types of error and taking some

> preventive action ... Incidental correction by contrast, is (1) confined to particular 'tokens', (2) only responsive (i.e. not leading to any preventive or pre-emptive action, (3) facilitative (i.e. regarded by learners as a part of getting on with the activity in hand, not as a separate objective and not as being more important than other aspects of the activity), and (4) transitory, (i.e. drawing attention to itself only for a moment - not for as long as systematic correction does).

The labeling here provides the following contrasts:

systematic vs incidental

long interruption vs transitory reference

explanation vs no explanation

exemplification vs no exemplification

preventive vs responsive

relating to types vs relating to tokens

a primary objective vs merely facilitative

In addition, the kind of learners' errors that would receive attention, according to Prabhu (1982, p.5) would be more to do with content than with linguistic accuracy, which would contrast with the structural approach, and give us the following contrast:

linguistic vs more content than linguistic

7.3 Formulation of Research Questions

7.3.1 Hypotheses relating to the congruence of CTP practice and CTP attitudes to error treatment

Although it is not known if any approach exists which is purely systematic in its treatment of error, at least the exercise of forming dichotomies is a movement

towards some kind of specificity about what natural control might be and how it is to be distinguished in classroom practice from planned control. It helps to make the issue researchable. From Prabhu's descriptions cited above and making use of the apparent dichotomies, the following hypotheses seem permissible:

1. In CTP lessons more content errors are treated than linguistic.
2. Treatment of linguistic error involves no explanation.
3. Treatment of linguistic error involves no exemplification.
4. Treatment of linguistic error involves no generalisation (i.e. no rules or types).

The notion of transitoriness is too loose to be researched. How brief and fleeting exchanges should be admits great latitude for interpretation, so no related hypothesis is proposed. Similarly the claim that CTP error treatment is simply facilitative of the task on hand is unfalsifiable. If it were shown that an utterance were contrary to the third hypothesis, it might be argued that on this occasion exemplification contributed to the task on hand, i.e. it was perceived by the teacher that it cleared up some doubt in the student's mind and allowed the task to proceed. However, an unfalsifiable claim cannot be used as an argument, or we are reduced to the kind of debate reported in section 7.1 above. Therefore, it seems justifiable to leave the

'facilitative' claim out of the present inquiry.

We have, then, 4 hypotheses to investigate.

7.3.2 Choice of a descriptive format to document types of error treatment

In addition to testing the 4 hypotheses, it would appear worth documenting in as much detail as possible just what kinds of treatment are given. This would require the selection of a checklist of treatment categories.

An examination of the literature revealed that there are a few models that have been devised for the expressed purpose of describing the corrective treatment of learner error in the classroom: Allwright (1975), Chaudron (1977) and Fanselow (1977b). The broader notion of 'repair' developed by Kasper (1986) out of the ethnographic work of Schegloff, Jefferson and Sacks (1977) does not seek particularly to categorise different types of teacher correction of pupil error, but is as much concerned with teacher self-repair and pupil self-repair; in view of this it will not influence our study. We are effectively left with the 3 models devised by Allwright, Chaudron and Fanselow.

Chaudron, in a later article (1986), notes that since 1977 no further models have appeared in the literature, and that this neglect is partly due to the influence of theories of second language acquisition that stress natural acquisitional processes (like the Natural

Approach of Krashen and Terrell 1983). Another proffered reason is that the connection between error treatment and success in learning has been perceived as one that is virtually impossible to establish. Nevertheless, Fanselow's framework was modified by Courchene (1980), and Chaudron's categories were adapted by both Nystrom (1983) and Salica (1981), suggesting that interest has not totally declined. Whatever the current level of interest, a study of the treatment of error is relevant to the present chapter.

Allwright's (1975) set of categories, which he calls "preliminary" (p.108), number 16; 7 'basic options' and 9 'possible features'. Fanselow (1977b) also has 16 categories only some of which approximate to Allwright's. Chaudron (1977) has identified 31. From our perspective, the number is irrelevant, but the categories should assist in the consideration of the 4 hypotheses of section 7.3.1, and also provide a rich description of the strategies used by CTP teachers.

A glance through the data indicated that rephrasing was an important strategy and Chaudron's breakdown of repetition into 4 types appeared potentially useful here. Also, the categories of 'altered question' and 'original question' were frequently reflected in samples of the lesson transcripts. Finally, since none of the categories used by Fanselow and Allwright were absent in Chaudron, the latter's framework was seen as a suitable point of

departure. (Chaudron's 1977 framework is presented in Appendix 19).

Chaudron's 31 categories were trialled on 5 CTP transcripts to test their adequacy. As a result of this probe, 11 categories were deleted since they were not coded. Also the descriptions of certain categories were slightly modified to better accommodate the data. The remaining categories and descriptions are now set forth, along with both linguistic (L) and content (C) examples from the transcripts. Line numbers are also given in brackets for easier reference, followed in the same brackets after a comma by a capital letter designating a particular lesson transcript.

IGNORE: Teacher ignores pupil's error and goes on to another topic.

Examples:

(L)(1,J)

T: I'll give you some tasks, right?

S: What is task?

T: I'll, I'll give. I'll tell you ... Now have you got rules? Where are your rules? Come on. Hurry up.

(L)(99,J)

S: The rural important because

T: Give me the first reason

ACCEPTANCE: Simple approving or accepting word (often as a sign of reception of the utterance, but teacher may

proceed to correct an error.

Examples:

(L)(115,G)

T: Who paid the money?

S: Who received the letters.

T: Yes, the man who received the letters paid the money.

(L)(228,J)

SS: Show in Calcutta.

T: Alright, I'll show it in Calcutta.

ATTENTION: An attention getter, like 'listen to me' or 'think'.

Examples:

(C)(262,T)

T: Is it complete?

S: Yes sir.

T: Look carefully.

(C)(100,U)

T: How many more kilometres does he travel?

S: Two hundred and ... nine, ninety.

T: Think carefully.

NEGATION: Teacher shows rejection of part or all of student utterance.

Examples:

(C)(467,A)

T: Educate. You know the spelling?

SS: e - j

T: No, there's no 'j' in it.

(C)(173,C)

T: Who's writing this letter?

S: Geetanjali.

T: Not Geetanjali.

S: Library member.

T: Not library member.

PROVIDE: Teacher provides the correct answer when Student has been unable to or when no response is offered.

Examples:

(C)(557,D)

T: I want the application form. For what?

S: Post office.

T: Application for ... for cancelling, application for cancelling the radio licence.

(L)(70,F)

S: In a

T: No, in <u>the</u>

REPETITION with NO CHANGE: Teacher repeats student utterance with no change of error nor omission of error.

Examples:

(C)(268,L)

S: Draw a square, draw a square with bottom of the midpoint.

T: Draw a square. [the answer is 'draw a circle']

(L)(243,J)

T: So, who is right now?

SS: Car driver. Car owner.

T: Car driver. [fails to say 'the car driver]

REPETITION with CHANGE and EMPHSAIS: Teacher repeats student utterance with no change of error, but emphasis locates or indicates fact of error.

Examples:

(L)(66,E)

S: Requested you. [uses wrong tense]

T: Requested?

(C)(554,D)

S: I read the application

T: I read the application [the answer is: 'I want the application]

REPETITION with CHANGE: Usually, teacher simply adds correction and continues to other topics.

Examples:

(L)(35,C)

S: Sheila 15 years old.

T: Sheila is fifteen years old.

(L)(116,I)

S: Don't know kerosene spelling.

T: You don't know the spelling of kerosene?

REPETITION with CHANGE and EMPHASIS: Teacher adds emphasis to stress location of error and its correct formulation.

Examples:

(L)(356,D)

S: Renew the licence

T: Renew my licence.

(L)(403,D)

S: Licence

T: Licences for two band radios.

EXPLANATION: Teacher provides information as to cause of error, possibly including a generalisation of the type of error.

Examples:

(L)(225,F)

S: Cow dung, cow dung was

T: No, not cow dung was. It is there. It's not was. It's not in the past. It's there now.

(C)(49,B)

S: Bags

T: They make bags, they make bags - all these come under 'handicrafts'; pots, bags, all that come under 'handicrafts', so we, this is not, I won't take 'bags' in this.

REPEAT: Teacher requests student to repeat utterance with the intention of having the student self-correct. This can only be distinguished from LOOP (below) by a subjective judgment and by the fact that the REPEAT prefigures a later attempt to elicit the correct answer.

Examples:

(C)(64,M)

T: Anything more to write?

SS: 5 ['5' is the wrong answer]

T: What?

SS: 5

(C)(28,Q)

T: Where is the long hand?

SS: 7, 7 ['7' is the wrong answer]

T: Hmm?

SS: 7

LOOP: Teacher honestly needs a replay of student utterance due to lack of clarity or certainty of its form. This can only be distinguished from REPEAT above by a subjective judgment and by the fact that the REPEAT prefigures a later attempt to elicit the correct answer.

Examples:

(L)(9,B)

S: I want weaving spelling, miss.

T: Sorry?

S: Weaving.

T: Weaving. Anyone in the class knows the spelling?

(L)(77,E)

S: /arli/

T: Once again?

S: /arli/

T: Early, yes.

PROMPT: Teacher uses a lead-in cue to get student to

repeat utterance, possibly at point of error; possible slight rising intonation.

Examples:

(L)(50,L)

S: Draw a two horizontal line

T: Draw ...

S: Draw two horizontal lines

(L)(390,A)

T: They collect water

S: Water, miss

S: In the well, in the well

T: They collect water ...

S: In the well

T: From. From the river or well.

CLUE: Teacher reaction provides student with isolation of type of error, or the nature of its immediate correction, without actually providing the correction. E.g. further examples of the same error type may be given.

Examples:

(C)(527,T)

T: From?

S: Jamali

SS: Jamali

T: Ramani wrote this letter

SS: Tiruchi

T: From Tiruchi, yes.

(L)(218,E)

T: Will you please send me a application form? [a student has said 'a application form']

S: No, no, sir, two.

T: Will you please send me a application form, a? A apple? Do we say a apple?

ORIGINAL QUESTION: Teacher repeats the original question that led to the incorrect response.

Examples:

(C)(54,M)

T: Is there anything more to write?

S: Yes [in fact, there is not]

T: Is there anything more to write?

(C)(162, D)

T: Who wrote the letter?

S: B.N. Rao

T: Who wrote the letter?

ALTERED QUESTION: Teacher alters original question syntactically but not semantically.

Examples:

(C)(56,M)

T: Is there anything more to write?

S: Yes [this is wrong]

T: Is it finished?

S: No [still wrong]

(C)(35,B)

T: Did I mention it in the talk or not? [she did]

S: No, miss.

T: Working in agriculture, was it mentioned? Was it mentioned in the talk?

QUESTIONs: Numerous ways of asking for a new response, but not just original or altered questions, i.e. when error occurs, a new line of questioning is taken up.

Examples:

(C)(7,O)

T: Is that alright? [referring to student's answer on the blackboard]

SS: No [in fact, it is alright]

T: What has he written?

(C)(80,J)

T: Will the police prosecute Chetyan?

S: Yes sir [this answer is wrong]

T: What is the meaning of 'prosecute'?

TRANSFER: Teacher asks another student or group of students to provide correction.

Examples:

(C)(34,F)

T: Is it alright now?

S: Yes miss [wrong]

T: All of you think it's alright now?

(C)(142,K)

T: The science lesson on Friday is just before history. Who will do that? Yes? [student comes to the board and writes 'scins'] ... Is that alright? [addresses class]

ACCEPTANCE*: Teacher shows approval of student utterance.

Examples:

(C)(36,H)

S: Sir, exchange [he wants to say: 'change' trains]

T: I have to exchange! Alright.

(L)(176,U)

T: Which is the shortest route?

S: Straight route [fails to say 'the straight route']

T: Straight route is the shortest route

VERIFICATION: Teacher attempts to make sure that the class has understood the correction.

Example: (only 1 instance was coded)

(C)(274,L)

T: What is this point now?

S: Bottom. Point of circle.

T: It's the midpoint of this circle. It's the midpoint of this circle. Correct? Is that correct? [having provided the answer, the teacher simply wants to see if the class follows his correction]

It will be clear from the descriptions and the examples of the above 20 categories that they overlap and that most of the error treatments require multiple coding. Thus one treatment might be associated with more than one category, e.g.

T: Who paid the money?

S: Who received the letter

T: Yes, the man who received the letter paid the money.

This was presented above as an example of ACCEPTANCE. It is, however, also an instance of REPETITION with CHANGE. Wherever multiple coding was applicable, every possible categorisation was counted, which is why there are many more treatments than errors.

7.3.3 Seeking explanation of the incidence of different types of error treatment.

In addition to the 4 hypotheses and description, it would be helpful from the perspective of external validity if the incidence of particular kinds of error or treatment could be explained.

An obvious candidate for investigation would be the differences between RTs and NRTs, since chapters 5 and 6 have shown that it is an important variable. However, our data do not include lessons given by RTs, so such an inquiry is impossible.

Recalling that the CTP was a developmental program, that is, it did not start life with a ready-made methodology, but evolved one over time through a process of trial-and-error, it seems quite feasible that error was treated differently in the early and later phases of the project. To make this distinction, it would be helpful to find a watershed, a period when the project methodology had become relatively stable, when it would

be unlikely to undergo any further major modification.

Barnes, visiting the project in March/April 1982, reports that the lessons are "impressively consistent" (1982, p.2). A letter to Prabhu sought his views on the question. He responded as follows (cf. chapter 5, section 5.4.1.3):

> The pattern of classroom activity was least settled in the first year of the project (June 1979 - April 1980), most settled in the last two years (June 1982 to April 1984) and was settling steadily in the second and third years (June 1980 to April 1981 and June 1981 to April 1982). Between those two years, teaching at a post-initial level was well-settled by about the middle of the second year (say, December 1980) while that at the zero level had a stable pattern by the middle of the third year (say, December 1981). (Prabhu, personal communication, May, 1986).

Coincidentally, none of our lessons took place in 1982, but were instead all either before or after that year. Thus, before or after 1982 would appear to be the appropriate division.

Another potentially important variable is the length of different lessons. If some of the recordings are shorter than others or incomplete, then this might have a bearing on our interpretation.

Yet another possible explanation of the incidence of different kinds of error treatment is that it is a matter of personal style, associated with some teachers but not with others. Of the 21 lessons at our disposal, 18 are taught by 2 teachers. It would be possible to compare them bearing in mind that with an N-size of 2, inference to a wider population is extremely tenuous.

Finally, a variable that demands attention is that of task-type. Is the incidence of error and type of error treatment dependent on the nature of the task? Again, this would influence any interpretation of treatment of errors in CTP classes.

Thus the third strand of the inquiry into the treatment of error comprises 4 elements: (i) location in time (pre-1982 / post-1982), (ii) length of recording, (iii) personal style, and (iv) task-type.

7.4 Method

In section 7.4, first of all the data will be described (7.4.1). This will be followed by a report of the procedures used to analyse the data (7.4.2).

7.4.1 The Data

(All transcripts of recordings used in the present study are presented in Appendix 20).

It was known that some recordings had been made of CTP lessons, but none of these was still in the hands of the director of the project, Dr. Prabhu. The Department of Applied Linguistics at the University of Edinburgh was able to produce copies of video recordings of 5 CTP lessons (lessons M, N, O, P, and Q in Appendix 20). 10 audio recordings were made available by the Department of Linguistics at the University of Lancaster and David Carroll, a former British Council KELT officer who had

helped in the development of the Bangalore project (these are lessons A to J in Appendix 20). In addition, the Department of Linguistics at the University of Lancaster made available 4 transcripts (lessons R to U in Appendix 20). A further 2 transcripts were sent by a CTP teacher (Teacher O)(lessons K and L in Appendix 20).

With reference to task-type, 5 of the lessons were lecturettes (A, B, F, G, and I); 4 involved letter-writing (C, D, E, and T); 4 revolved around timetables (H, K, N, and R); 3 were based on drawing figures (L, M, and O); 2 focused on town maps (P and S); 1 on distances (U); 1 on telling the time (Q); and one on short narratives plus questions (J). Lecturettes were long narrative descriptions (e.g. of the development of the current postal system) broken periodically by oral comprehension questions. Letter-writing consisted of the teacher eliciting from the students the necessary language for a coherent letter on a predetermined topic to be composed on the blackboard - initially, students were presented with scripts of the letter containing error; the students were to improve these scripts. Time-tables required logical reasoning in order to complete empty cells. Figure drawing demanded that students listen to instructions and draw what they are told to draw. Town maps require students to follow directions. Lessons about distances focus on the ability to make simple calculations, totting up or subtracting distances between towns. Telling the time is self explanatory except that

inference is required to answer many of the questions.

Of the 21 lessons, 10 were taught by teacher O (A, B, F, I, K, L, N, O, S and U), 8 by the project director (D, E, G, M, P, Q, R and T), 2 by teacher I (C and J), and 1 by an unidentified teacher (H). (We know, then, that at least 20 of the lessons were taught by NRTs, thus precluding a comparison with RTs). The only sensible comparison can be between teacher O and the project director.

As for the length of the lessons, some were clearly incomplete. Also, recordings of the task phase of each lesson were inaudible as the teacher would go around the class checking the work of individuals. Therefore, it was only possible to investigate the pre-task phase of each lesson. Even here, the range varied from 6 minutes to 39 minutes, with a mean of 22.67 minutes and a standard deviation of 8.67 minutes. This analysis includes only the recorded lessons; for the 6 lessons where only transcipts were available, the duration is not known.

Lessons A to J and R to U all dated from February and March 1981 (14 lessons). Lessons K to Q were from 1983 and 1984 (7 lessons).

All of the information about the data is summarised in Table 7.1, following.

Table 7.1

Summary information on available lesson transcripts

Lesson	Year	* A/V/T	** Length	Teacher	Task-type
A	1981	A	39	O	Lecturette
B	1981	A	27	O	Lecturette
C	1981	A	12	I	Letter-writing
D	1981	A	33	Director	Letter-writing
E	1981	A	24	Director	Letter-writing
F	1981	A	20	O	Lecturette
G	1981	A	30	Director	Lecturette
H	1981	A	6	Unknown	Train Timetable
I	1981	A	20	O	Lecturette
J	1981	A	13	I	Short narrative
K	1983	T	?	O	School Timetable
L	1984	T	?	O	Figure drawing
M	1984	V	16	Director	Figure drawing
N	1984	V	18	O	School Timetable
O	1984	V	25	O	Figure drawing
P	1984	V	24	Director	Town maps
Q	1984	V	33	Director	Telling the time
R	1981	T	?	Director	Train Timetable
S	1981	T	?	O	Town maps
T	1981	T	?	Director	Letter-writing
U	1981	T	?	O	Distances

* 'A' refers to Audio, 'V' to Video, and 'T' to Transcript only. ** Length is in minutes.

7.4.2 Procedures

7.4.2.1 Transcription Conventions

The following transcription conventions were used:

T = Teacher

S = Individual Student

SS = More than one student speaking at once, possibly the whole class

(S
(= Simultaneous speech
(T

(Anything in brackets is part of the commentary, e.g. 'student comes to the board')

X = an incomprehensible utterance, apparently of word length

XX = an incomprehensible utterance, probably of phrase length

XXX = an incomprehensible utterance, beyond phrase length

Words are sometimes separated by three dots: ... This refers to a pause of more than 2 seconds and less than 10 seconds.

When there are 6 dots: ..... this refers to a pause of more than 10 seconds and less than 30 seconds.

(PAUSE) = a break in speech of 30 seconds or more.

Punctuation is normal, including full-stops, commas, question marks and capital letters.

Emphasis is conveyed through <u>underlining</u>.

The choice of these conventions approximately follows those used by 4 M.A. students at the University of Lancaster who transcribed and analysed 4 CTP lessons (under the supervision of Dick Allwright of the Department of Linguistics). (These lessons are used in the present study). Similar conventions were also used by teacher O in her transcription of 2 of the lessons used in the present study. So, in order that there should be some measure of uniformity between the 6 lessons that had already been transcribed and the 15 that remained, the above conventions were adhered to.

It might reasonably be argued that the use of punctuation imposes a judgment on the text. Clearly this is true, but such judgments were essential to our study. It was necessary to know what could be classified as questions, for example, so the judgment was made at the transcription stage, and question marks inserted. Although for certain kinds of inquiry, it would be preferable to reserve judgment until a later stage, so that one may always return to a relatively unreconstructed text, for the purposes of identifying types of error and types of treatment such judgments as were made appeared to be appropriate to the study.

7.4.2.2 Steps in the Inquiry

The first step was to identify errors. The three following strategies were adopted: (i) a judgment was made concerning the linguistic accuracy of a student's

utterance, (ii) a judgment was made as to the accuracy of a student's response to a teacher's question in terms of content (i.e. was the solution to a problem correct or incorrect?), and (iii) if a teacher clearly disapproves of a student's utterance, the utterance was considered to contain error (in practice, this meant that we were alerted to the likelihood of error and sought to identify it).

The second step was to categorise the error as either linguistic or content. By 'linguistic' error is meant morphosyntactic or phonological error. In the event, only 4 phonological errors were noted; in these few cases, the teacher failed to understand the students' utterances. This failure to understand was used as a criterion. A stricter criterion could have meant classifying perhaps almost every single utterance as phonologically deficient in some way, which would gain nothing.

By 'content' error is meant any response by a student to a teacher's question which isunsatisfactory in terms of its propositional content. Thus, if a student answers a question that was not asked, or simply answers the right question wrongly, a content error was coded. For example, if the teacher's question requires a calculation to which the answer is '6' and a student answers '7', a content error has been made which is irrespective of the linguistic accuracy of the utterance. (4 lexical

errors were noted and were included as content errors)

Any error which was both linguistic and content was classified as both. All errors are listed for each lesson in Appendix 21.

The third step was to calculate the percentages of linguistic and content errors that were treated (to permit a judgment with reference to hypothesis 1).

Following this, treatments were identified in terms of the 20 categories listed and exemplified in section 7.3.2 above. This proved to be a very demanding process as it required a great deal of working back and forth between codings given at different sittings and for different lessons to try to ensure a reasonable level of consistency. Since this analysis was carried out only by the author, some form of self-checking protocol seemed appropriate. Thus, after all of the lessons had been painstakingly coded, 3 lessons were recoded months later. The agreement was only 71%. It was found that the difficulties of handling the data consistently had not been overcome by the early efforts. It might be speculated that another analyst, not knowing what subtleties had influenced the author and not having his memories, would have agreed with our coding at a considerably lower level than 71%. The difficulties of such analysis are well attested to in Chaudron (1977), Allwright (1975) and Fanselow (1977b)(see also chapter 2, section 2.3.2.1), but the lack of intracoder consistency needs to be stressed, as intercoder reliability would

probably be somewhat lower.

Once the error correction strategies had been coded, the transcripts were scanned for treatment of linguistic error which involved explanation, exemplification or generalisation (in order to test hypotheses 2, 3, and 4).

Finally, the incidence of various strategies was calculated (for the descriptive purposes outlined in section 7.3.2) and the analyses were arranged for examination of the potentially explanatory variables discussed in section 7.3.3.

7.5 Results

7.5.1 Treatment of Linguistic and Content Errors

Once all errors had been identified and classified as either linguistic or content, it was possible to calculate the percentages of linguistic and content errors that were treated. The total number of error treatments for the 21 lessons is 926 ($\bar{x}$ = 44 per lesson); the total number of content error treatments is 599 ($\bar{x}$ = 29 per lesson); and the total number of linguistic error treatments is 327 ($\bar{x}$ = 16 per lesson).

'Treatment' includes the categories IGNORE, ACCEPTANCE, and ACCEPTANCE*. If, however, we deduct these categories (which are really non-treatments) from the number of treatments, a more accurate picture emerges of the comparative attention given to linguistic and content errors. We find that only 193 out of 327 linguistic

errors were treated (65%), while 529 out of 599 content errors were treated (88%). On a 2 X 2 chi-square test, p = <.0001, which indicates that the first hypothesis was correct: In CTP lessons, more content errors are treated than linguistic.

7.5.2 Explanation, Exemplification and Generalisation

On completion of the categorisation of error correction types according to the criteria put forward in section 7.3.2 above, the treatments were scanned for instances of the use of explanation, exemplification and generalisation by the teacher when a student had made a linguistic error. This information would contribute to the acceptance or rejection of hypotheses 2, 3, and 4 (see section 7.3.1).

The categories in the framework that were intended to track these strategies are EXPLANATION and CLUE. EXPLANATION includes both giving information about the cause of the error and making a generalisation about the type the error represents. CLUE includes the use of further examples of the same error type.

There were very few examples of EXPLANATION in the treatment of linguistic error (N = 4) when compared with the far greater frequency of this strategy to deal with content errors (N = 50).

3 of the 4 events occurred in one exchange, beginning on line 66 in lesson E:

S: Requested you

T: Requested? When?

S: (Request
(

T: (Request. He's requesting now [1]. If he requested yesterday, then we can say 'requested' [2].

S: Yes sir.

T: Now he's requesting the post office [3].

The 3 instances of EXPLANATION are numbered in square brackets. The first provides the student with information about the cause of the error - there is something wrong with the marking of time (the location of which error has already been highlighted on the second line of the exchange). The second instance goes even further; a generalisation is made. In the third instance, the cause of the error is stressed.

A fourth instance of EXPLANATION comes in an exchange beginning on line 225 of lesson F:

S: Cow dung, cow dung was

T: No, not cow dung was. It is there. It's not was. It's not in the past. It's there now.

The teacher's response quite clearly focuses on the cause of the error and generalises to the extent of invoking the concept of the 'past'.

Turning to the question of exemplification, it was found that CLUE was coded 6 times for linguistic error, but closer inspection revealed that only one of these was

a matter of exemplification. The exchange involving this instance begins on line 209 of lesson E:

T: You please send me?

SS: Application form.

T: Application form. Do we want anything here? Application, application. Will you please send me a application form?

S: No, no, sir. Two.

T: Will you please send me a application form? Application form, a? ... a apple? Do we say a apple?

SS: (Laughing).

T: Yes? You want one more letter here.

S: An application.

T: Right, an application.

The point where exemplification of the type of error comes in is where the teacher says "a? ... a apple? Do we say a apple?".

Considering the above 3 exchanges, it seems clear that the nature of the focus on form is such as to contravene Prabhu's statements about the ways in which error should be handled to be consistent with CTP principles. Explanation, generalisation, and exemplification are all used, although rarely, in ways that do not conform with hypotheses 2, 3, and 4 put forward in section 7.3.1. If the strict wording of those hypotheses is adhered to, ("no explanation", "no

generalisation", "no exemplification"), then the hypotheses are rejected.

On the other hand, it is quite clear that the incidence of such 'violations' is very slight. Also, all except one of the exchanges occur in lesson E. However, lesson E was taught by the project director, and if the instigator and chief proponent of the CTP could slip into a focus on form, then it is hardly surprising that other CTP teachers did (see chapter 5). From the point of view of external validity, it is important to note that not only RTs reverted to more familiar classroom practice; further, it is not an unreasonable inference that if the director could focus on form occasionally, then the degree to which this was true for the RTs is likely to have been considerable.

7.5.3 A Descriptive Account of the Treatment of Error

Applying the criteria laid out in section 7.3.2 above, the incidence of different treatment types for linguistic and content errors is as displayed in Table 7.2 below.

Table 7.2

A Descriptive Summary of Error Correction

Treatment	Ling.	%	%Total	Cont.	%	%Total
IGNORE	61	57	18.7	46	43	7.7
ACCEPTANCE	36	61	11.0	23	39	3.8
ATTENTION	0	0	0.0	4	100	0.7
NEGATION	9	8	2.8	101	92	16.9
PROVIDE	9	20	2.8	37	80	6.2
REP/NO CH.	12	50	3.7	12	50	2.0
REP/NO CH + EMPH.	6	46	1.8	7	54	1.2
REP/CH.	116	97	35.5	4	3	0.7
REP/CH + EMPH.	7	78	2.1	2	22	0.3
EXPLANATION	4	7	1.2	50	93	8.4
REPEAT	1	25	0.3	3	75	0.5
LOOP	4	100	1.2	0	0	0.0
PROMPT	13	29	4.0	32	71	5.3
CLUE	6	12	1.8	44	88	7.4
ORIGINAL QUESTION	8	11	2.5	68	89	11.4
ALTERED QUESTION	10	16	3.1	54	84	9.0
QUESTIONs	3	5	0.9	62	95	10.4
TRANSFER	3	6	0.9	48	94	8.0
ACCEPTANCE*	19	95	5.8	1	5	0.2
VERIFICATION	0	0	0.0	1	100	0.2

Note: '%' refers to the % of a category between Ling. and Cont.; '%Total' refers to the % of a category within Ling. or Cont. Due to rounding off, column totals do not add up to 100 exactly.

With regard to linguistic error treatment, Table 7.2 indicates that by far the most common strategy is REPETITION with CHANGE (N = 116), which is to say, when a student makes a linguistic error, almost 36% of the teachers' correction strategies consist of repeating the student's utterance in an accurate form and moving on. This is entirely consistent with the CTP perceptions of appropriate error treatment.

Similarly appropriate is simply to ignore the error. IGNORE figures largely in CTP teachers' strategies - 61 instances or nearly 19% of the total range of strategies. Whatever may be thought of the pedagogic value of ACCEPTANCE and ACCEPTANCE*, they are within the CTP frame of acceptable treatments, occuring 36 (11%) and 19 (5.8%) times, respectively. Altogether, our data show that 222 out of 327 (71%) linguistic error treatments entail either a simple, unstressed rephrasing or a willingness to let the error pass altogether. When this is contrasted with the incidence of the same strategies in relation to content errors, (74 out of 599, or 12.4%), there is obviously a massive difference in general approach.

Apart from those mentioned, the rest of the strategies used to correct linguistic error are few in number and more or less neutral with regard to CTP policy.

By contrast, content errors evoked a wider range of treatments. The incidence of different types is more

evenly spread. This may reflect more sustained attempts to bring about in the learner a preoccupation with solving a problem correctly. When one form of treatment fails, others are tried.

The overall picture provided by Table 7.2 is fairly clear. A few major strategies dominate in correction of linguistic error, and these reflect a general willingness to allow errors to pass with a simple rephrasing, without comment or even with apparent acceptance. This is in keeping with expressed CTP attitudes to error correction. Content error, on the other hand, receives far more sustained and varied treatment, possibly indicating an emphasis on solving the problem at hand. However, this overall description hides the contribution of potentially relevant factors, such as location in time, length, teacher style and task-type. The following section addresses these factors.

7.5.4 Variables Contributing to Incidence of Treatment Types

In this section, we examine the influence of the following 4 variables: location in time (pre-1982 / post-1982) (section 7.5.4.1), length of pre-task (section 7.5.4.2), teacher style (section 7.5.4.3) and task-type (section 7.5.4.4).

7.5.4.1 Location in Time (Pre-1982 / Post-1982)

As elaborated in section 7.3.3, one possible

explanation of the spread and the frequency of types of corrective treatment is that the CTP methodology had not achieved the relative stability of later years until about 1982. Of the lessons available to this study, 14 took place in 1981 (A to J and R to U) and 7 in 1983/4 (K to Q). Table 7.3 details the frequency of linguistic and content errors and the percentages of correction (i.e. deducting the IGNORE, ACCEPTANCE and ACCEPTANCE* strategies) for the 2 sets of lessons.

Table 7.3

Frequency of Errors and Corrections Pre- and Post-1982

	Pre-1982 (N = 14)		Post-1982 (N = 7)	
	Ling.	Cont.	Ling.	Cont.
No. of Errors	243	247	36	136
$\bar{x}$ no. of Errors	17.36	17.64	5.14	19.43
No. of Corrections	162	206	31	126
% Corrected	67	83	86	93

It can be seen from Table 7.3 that there is a slightly higher incidence of content error in the post-1982 lessons (an average of 19.43 compared with 17.64). There is also a significantly higher rate of correction (using chi-square, $p = >.05$ with 1 degree of freedom). With regard to linguistic error treatment, too, there is

a significantly higher percentage (using chi-square, p = >.05 with 1 degree of freedom). However, the most salient difference is that in the pre-1982 lessons, the mean number of linguistic errors per lesson is 17.36; in the post-1982 lessons, by contrast, this mean has fallen to 5.14. The immediate question that emerges is 'why are so many fewer errors being made in the later lessons?'. It was possible that length of pre-task, teacher style or task-type could account for this very substantial drop.

7.5.4.2 Length of the Pre-task

Of the 14 pre-1982 lessons, 4 were made available to the author in transcribed form, so the timings are only known for the remaining 10. Similarly, in relation to the post-1982 lessons, 2 were already transcribed and the timings are only known for the remaining 5 (see Table 7.1). For the 10 early lessons, the mean duration of the pre-task phase was 22.4 minutes, compared with 23.2 minutes for the 5 late lessons. Duration was practically equivalent and therefore could not be judged likely to have had a bearing on the observed pre- and post-1982 differences.

7.5.4.3 Teacher Style

Since the majority of the 21 lessons were taught by teacher O (10 lessons) and the Director (8 lessons), only these were examined. Table 7.4 shows that a very similar rate of error occurs in both teachers' lessons, and that

the percentage of errors corrected is also similar.

Table 7.4

Frequency of Error and Correction

	Teacher O		Director	
	Ling.	Cont.	Ling.	Cont.
No. of Errors	112	178	109	181
$\bar{x}$ No. of Errors	14.00	17.80	13.63	22.63
No. Corrected	82	159	74	152
% Corrected	73	89	68	84

Allowing that with such a small sample of teachers, there is very little to go on, there is no evidence to suggest that teacher style might be a variable that can explain the differences between early and late lessons.

7.5.4.4 Task-Type

Although the mean number of linguistic errors for the later lessons is 5.14, closer inspection of the data reveals that 25 of the 36 errors occurred in one lesson. This lesson is different in kind from the others in that instead of just of just receiving instructions, students give them too. Teacher O, who taught the lesson, explains in her account (Appendix 7) that the intention was to increase learners' productive ability in English; she points out that this was atypical of CTP teaching. If this atypical lesson is discounted in the pre- and post-

1982 analysis, the later lessons average only 1.83 errors which, compared with the 17.36 of the earlier lessons, is a startling difference. Thus, on a first examination of task-type, it would appear that there are barely any linguistic errors in the later lessons.

Barely any errors require barely any treatment. So the question now becomes 'what is the association between task-type and the incidence of linguistic error?' And once this is answered, it would be germane to ask whether task-types that have little association with linguistic error occur more often or solely in the later period. In other words, did the CTP phase out tasks that appeared to draw student responses that might contain morphosyntactic error?

Table 7.5 indicates the mean incidence of linguistic error that is associated with each task-type (not including teacher O's figure-drawing lesson that contained student-student instructions, because of its atypicality):

Table 7.5

Linguistic Error and Task-Type

Short Narrative	30.00
Letter-writing	27.75
Lecturettes	15.80
Distances	6.00
Timetables	5.50
Town Maps	4.00
Telling the Time	4.00
Figures	1.00

Table 7.5 arranges the task-types according to pre-1982, post-1982 or both periods, and includes in brackets the frequency of their occurrence in each period.

Table 7.6

Task, Error and Location in Time

Pre-1982	Pre- and Post-1982	Post-1982
Lecturette (5)	Timetables (4)	Figure-draw. (3)
Letter-writ. (4)	Town Maps (2)	Telling Time (1)
Distances (1)		
Short Narratives (1)		

If Tables 7.5 and 7.6 are considered in conjunction, it can be seen that the tasks which are most associated with linguistic error occur only in the pre-1982 period and that those which occur in both periods and in the post-1982 period only are characterised by a low frequency of such error.

It would appear, then, that task-type has a bearing on the incidence of linguistic error, and also that the CTP phased out those tasks which tended to induce grammatically inaccurate responses from students.

Examining the data more closely, it emerges that 198 out of the 200 linguistic and content errors in the Letter-writing tasks are verbal, whereas in the Figure-drawing tasks, only 12 out of 46 are verbal (and 10 of them are simply wrong answers to yes/no questions or wrong letters or numbers called out).

In the Lecturettes, the teacher talks at length on a theme and then asks a series of comprehension questions which prompt verbal responses usually of phrase length, but sometimes longer. In Letter-writing, students are presented with scripts containing error which they are required to improve, so that language itself becomes the focus of the task and greater demands are placed on students' verbal productive capacities. The figure-drawing lessons, by contrast, call for mainly physical responses (i.e. writing or drawing on the blackboard); thus, whether the response is right or wrong, there is

very little possibility of the student making phonological or morphosyntactic errors.

Aware that our sample of 21 lessons could well be unrepresentative, the RIE Newsletters and Bulletins, the cyclostyled lesson reports, and Prabhu's writings were examined for further clues. There is no mention atall of the Lecturettes and no post-1982 mention of Letter-writing. All of the tasks which demand little verbal production, on the other hand, receive frequent mention. Figure-drawing, Telling the Time and Timetables are referred to as 'beginners' tasks' (Prabhu 1987, p.37-39), presumably because they focus on receptive skills, but it may be that as the CTP matured, there grew a tendency to protect learners from the risk of linguistic error attendant upon production beyond the beginner level, and that tasks such as Lecturettes and Letter-writing were dispensed with. Certainly, they are not included in the long list of task-types presented in Prabhu (1987, p.138-143).

The story that seems to be taking shape is that the CTP started out with a range of tasks that permitted or encouraged not only receptive but also productive skills; that as the project methodology settled, it moved towards delayed production and physical rather than verbal responses, probably not just at the beginners level, as teacher O's perception that the CTP did not attend sufficiently to production (Chapter 5 and Appendix 7) attests. But it is stressed that this cannot be

considered the full story; a sample of 21 lessons, although endorsed by all available evidence, is still only capable of supporting tentative statements.

7.6 Summary and Discussion

In this chapter, the aim has been to obtain a clearer perspective on the distinction between 'planned' and 'natural' language control in CTP practice. The relative specificity of Prabhu's (1982, and forthcoming) discussion of the CTP treatment of student error provided a means of getting at one aspect of the distinction.

21 CTP lesson transcripts were analysed (i) for the relative incidence of content and linguistic error treatments, (ii) to determine whether the statements against the use of generalisation, exemplification and explanation were strictly observed, (iii) to provide a descriptive account of CTP error treatment and (iv) to explore the possibility that certain variables might explain the incidence of different treatment types.

It was found that content error was more likely to be treated than linguistic error, which is consonant with the CTP focus on meaning rather than form.

For the descriptive account, a modified form of Chaudron's (1977) framework was used: it transpired that the majority of treatments for linguistic error involved minimal intervention (merely rephrasing, i.e. REPETITION with CHANGE) or no intervention (IGNORE, ACCEPTANCE and

ACCEPTANCE*). Content errors, by contrast, were treated in a wide variety of ways, indicating that more sustained attempts were made to secure the correct answers to problems. The descriptive account tended to support the stated CTP attitudes to error correction.

The stipulation that explanation, generalisation and exemplification be avoided was not strictly observed; the fact that not only RTs but also NRTs could slip into a focus on form was noted.

Examining variables that might contribute to treatment type, a pre-1982 / post-1982 distinction was made. The most salient finding here was that in looking for a relationship between location in time and error treatment, there was very little error treatment. This was due to the simple fact that there was very little error. Excluding one atypical lesson, there was only an average of 1.83 errors per lesson in the later period compared with 17.36 in the earlier period. This could not be explained by the length of the recordings, which was practically the same for both periods. Nor could teacher style provide a clue; very similar rates of error and treatment were found in the lessons of both teachers examined.

The best explanation seemed to be that certain tasks were more highly associated with linguistic error than others and that these tasks occurred only in the earlier period. Considering the nature of the tasks, it emerged that the tasks associated with little linguistic error

called for physical responses and extremely limited verbal responses. The earlier tasks required greater verbal production. Early tasks like Lecturettes and Letter-writing are not even mentioned in Prabhu's (1987, p.138-143) list of task-types.

The implication of these findings appears to be that task-types that called for production were phased out and only those that stressed reception were retained. When this interpretation is considered along with teacher O's testimony that "generally our tasks were cognitively very challenging, but linguistically did not make adequate demands on the learners' productive abilities" (Appendix 7), confidence in the interpretation is increased. And if it is correct, that is, if task-type is selected to curtail student production, then it might be reasonably argued that this is a form, not of 'natural', but of 'planned' language control.

Prabhu has argued from the early days of the project that CTP practice is based on a belief that reception comes before production and that a period of incubation is to be reckoned with, i.e. production will occur when it is ready (1980a, p.21). All of the lessons in our study post-date these statements and yet the differences between the 1981 set and the 1983/4 set are clear. But in any case, there is presumably a difference between 'forcing' production (which Prabhu disapproves of) and 'permitting' production (which is markedly truer of the

1981 lessons than the later ones).

Bearing in mind the limitations of our data and analysis, the following tentative findings are suggested by the present study. On the whole, the treatment of error conforms with the fairly precise attitudes stated by Prabhu; however, the distinct possibility was raised that while the treatment of error may be consonant with natural language control, the virtual exclusion of error itself in the later years of the project (through selection of task-type) may be a form of planned language control. From the perspective of external validity, i.e. extrapolating to other circumstances, this would be an important consideration.

CHAPTER 8

CONCLUSIONS

8. Conclusions

This thesis has attempted to provide an evaluation of the Communicational Teaching Project. Its principal aim has been to analyse, interpret and present data in such a way as to facilitate extrapolation by interested parties from the circumstances that were investigated to other settings.

This perspective derived from the reviews of the literature of chapters 1 and 2. Chapter 1 examines the language teaching method studies to seek guidance in the conduct of the Bangalore evaluation. An analysis of 51 inquiries shows that major problems that have afflicted the field are (i) the lack of program-fair measuring strategies, and (ii) the lack of adequate monitoring of classroom process. The purpose of the analysis was not to imply a value judgment about the merit of the studies; rather, it was to ensure that the present study was sufficiently aware of the major difficulties facing an evaluation.

Chapter 2 turns to the educational and psychological literatures for assistance, in view of their far greater attention to program evaluation. Firstly, an overview of the recent history of the development of evaluation is presented, along with an outline of the vast range of so-called models that have been devised and trialled in the past 20 years. It emerges that there are a number of approaches to evaluation, any one, or combin-

ation, of which may be deployed to answer particular needs. No one form of inquiry is suitable for all evaluations in all circumstances.

The issue of program-fair testing is considered in detail, and it is noted that no strategies adopted to date have contrived to ensure that testing can be free of bias toward one of the comparison programs; neither the use of standardised tests, special tests for each program, common/unique objectives, nor an appeal to consensus. It was considered, however, that the literature was clear on two points: first, the use of standardised tests is probably misguided and, second, that at the very least, program-specific tests should be devised for each program. (Chapter 4 tried to build on this by including program-neutral tests in addition to the program-specific tests).

Chapter 2 also raises the question of monitoring program implementation. Here, the educational and psychological literatures are especially informative. There is unanimity that implementation does indeed need to be thoroughly documented. Many approaches are suggested, but the most widely used appears to be systematic observation, either real-time coding or working from recordings.

The review proposes that evaluation should accord priority to practice and that theory is only a secondary concern. It is shown that the disappointment with field studies, such as Scherer and Wertheimer's (1964) inquiry,

has led many researchers (e.g. Freedman 1971) to opt for laboratory research. It is argued that since laboratory research can only indicate what occurred in a controlled environment, and is devoid of generalisability, it is not able to suggest what is likely to happen in particular schools. Field studies, on the other hand, can be relevant to program development, adoption and adaptation. A field study can arrange to be user-oriented insofar as it can present data analysis and interpretation so as to facilitate extrapolation by researchers, teachers and administrators. It can therefore be relevant.

Trading off elegance for relevance (or internal validity for external validity) is regarded as a justifiable approach to evaluation because programs (including methods) are too vague to be submitted for targeted inquiry in which it is known precisely what manipulations are of interest.

Chapter 2 above all stresses that the principal goal is to be relevant by enabling readers of evaluation reports to extrapolate, and that this may be a more prudent aspiration than to expect to contribute to a theory of language learning.

In Chapter 3, a description of the development of the CTP, its principles and methodology is presented. In addition, the literature that relates to the project is critically reviewed. It is clear that while the CTP has produced a great deal of interest, it has also provoked considerable controversy, particularly with regard to its

apparent lack of attention to evidence in the form of evaluation, its lack of learner-learner interaction, its reliance on referential language, and the possibility of a hidden language syllabus.

The interest in the project largely stemmed from a perception that it had the potential to inform about a prominent current model of language learning (that stresses unconscious processes). It is, however, beyond the scope of the present evaluation to assist in the understanding of a learning theory, as the attitude to evaluation emerging from Chapter 2 makes clear.

Chapters 4, 5, 6, and 7 attempt to evaluate the project from different perspectives. Chapter 4 compares the effect of CTP treatment with the structure-based program that is prevalent in South India. Chapter 5 reports a study in which the CTP teachers write detailed accounts of their experience on the project, and these accounts are subjected to an analysis to determine the levels of implementation relating to each of the teachers. Chapter 6 considers the concerns felt by teachers at different stages of their association with the CTP. Finally, Chapter 7 investigates the congruence of the project's stated attitudes towards the treatment of error (an element of language control) and the treatment of error observed in 21 lesson transcripts.

The testing component of chapter 4 finds that there is a difference in the language abilities resulting from

CTP teaching and those arising from structure-based teaching, and that grammar acquired during CTP training is more readily available for deployment than the grammar learned during structure-based training. The results also suggest that non-syllabus-based grammar can be learned without a focus on form (though not that it is best learned that way). However, these results are heavily qualified by a number of considerations relating to internal validity.

Principally, there was no systematic observation of comparison classes; thus we do not know enough about what actually happened in the classroom. Also, the groups, though not assigned to experimental and control treatments on a differential basis, were not randomly assigned; thus, there may have been initial differences in the constitution of the groups (no baseline measures were taken). Moreover, there was attrition from the classes and absence form the evaluation tests (these issues were extreme in the case of the T. Nagar school which was therefore dropped from the analysis).

With regard to external validity, the fact that most of the teachers were not regular teachers in the schools but more highly qualified personnel drafted in for the operation raises scruples about whether or not the results were largely due to them and not to the treatment. In addition, the likelihood of Hawthorne and novelty effects were noted.

Finally, a factor analysis of the results suggested

that while the relationships between the tests were by and large the predicted ones, there may have been some bias in two of the ostensibly program-neutral tests towards the CTP groups.

The findings of chapter 5 were that only two teachers apparently implemented the project and modified it according to their own perceptions. Most teachers seemed concerned to keep to perceived guidelines in spite of any wishes to, for instance, point out a structural distinction or use groupwork. Teachers who appeared to implement the least tended to be the regular teachers who, it was judged by the other teachers (and by visiting ELT experts), would be inclined to revert to the familiar focus on form, partly because of disciplinary problems associated with the sightly freer CTP lessons, partly because their English was inadequate, but also because the project teaching placed excessive demands on their time.

Chapter 6 provides a detailed profile of each teacher's concerns and also a peak concerns interpretation at each of three times (before project teaching, shortly after starting CTP teaching, and towards the end of CTP teaching). It is found that modifying the CTP remains a minor concern at all three times, that concerns with self and with coping tend to fade before concerns with impact emerge, that RTs are far more concerned overall at all three times than NRTs, that

concerns with self are more pronounced for NRTs than for RTs (who are more concerned with the effects of the CTP), that NRTs generally are concerned to conform to a CTP 'model' rather than to adapt and that RTs are not in a position to think beyond what they are still struggling to come to terms with.

Chapter 7 found that more content than linguistic errors were corrected in CTP lessons (consistent with stated CTP attitudes), but that contrary to these attitudes, explanation, exemplification and generalisation were used as corrective treatments, albeit rarely (in our data). The application of a modified version of Chaudron's (1977) descriptive framework indicates that stated attitudes to error treatment were on the whole adhered to (but no RTs were represented in our data); that is, linguistic errors were mostly dealt with either by rephrasing, ignoring or accepting, while content errors received a wider range of more sustained treatments.

It was also noted that linguistic errors hardly occurred in the post-1982 lessons (when the CTP methodology had stabilised) in comparison with the pre-1982 period. This could not be accounted for either by the length of the pre-task recordings or by teacher, but it seemed that task-type could account for the discrepancy. Inspection of task-type and the incidence of linguistic error in the two periods showed that most errors were associated with a few task-types which appear

to have been phased out in the later years of the project. It is suggested that the apparent attempt to curtail student production (and thus the potential for linguistic error) by the selection of task-type may be considered a form of planned language control; that in spite of avowed CTP intentions not to force production at the early stages of learning, the choice of task-type seems to have the effect of barely allowing it.

Findings from Chapter 4 suggest that although there did indeed appear to be a difference that favoured the experimental students, there were many potential explanations of this that had nothing to do with CTP training. One of the most obvious threats was that highly-qualified, highly-motivated personnel (NRTs) were drawn into project teaching; perhaps with regular teachers (RTs) the same differences may not have been realised. Chapters 5 and 6 indicate that the NRTs were aware of being pioneers and part of a close-knit team, that they tended to implement the CTP with greater fidelity, and that they were more anxious about the effects on students than they were about coping. RTs, by contrast, became project teachers when the CTP arrived at their schools, they had difficulty implementing the project, and they were concerned about their personal adequacy for the task. Thus, the findings of Chapters 5 and 6 tend to strengthen the doubts about regular teachers raised in Chapter 4.

Chapter 7 suggests that NRTs (even the director) were also prone to slip into a focus on form. When this is considered in conjunction with the views of visitors to the project and those of CTP teachers that RTs tended to revert to more familiar structure-based procedures, (including the admission by one teacher that she trained the weaker students in grammar separately), and that task-types permitting student production disappeared in the latter stages of the project's life, it would appear that there may have been a considerable level of planned language control in CTP classrooms.

Like the testing component of the evaluation (Chapter 4), Chapters 5, 6, and 7 are considerably constrained in the certainty of their conclusions. They all require analyses that are based on subjective judgment; Chapters 5 and 6 ask CTP teachers to retrospect which may well be even more open to distortion than introspection; in Chapter 6, the reliability and the validity of the questionnaire were not satisfactorily established; in Chapter 7, we are dealing with a small sample of lessons and an analytical framework that has only moderate intracoder reliability over time. The tentative nature of all of our findings is fully documented throughout the present study and is only recalled here for purposes of emphasis.

However, while the study by its very nature lacks experimental rigour, it aspires to be a disciplined inquiry, insofar as it attempts to make every step in the

compression and rearrangement of data transparent so that readers may follow the procedures closely and make their own judgments as to the credibility of each finding. This is perhaps desirable for all naturalistic forms of inquiry, but it is essential for an evaluative study which claims to arrange data to facilitate extrapolation by interested parties. For that, a minimum condition is that they should be able to make judgments of credibility.

Although the kinds of extrapolations that may be made are limitless, the closer the circumstances of the Bangalore study are to the circumstances one wishes to extrapolate to, the greater the likelihood of accurate judgment. The more dissimilar the two situations on a variety of parameters, the more tenuous and abstract the process becomes. Thus, extrapolation would be stronger to another South Indian mission school with large classes of pupils aged between 8 and 13 than it would be to an English Grammar school.

The extrapolator would perhaps wish to consider whether the teachers in his or her situation would view the project ideas with enthusiasm, whether they have adequate English, how pupils would react to a course with no linguistic syllabus; it might be considered whether, in view of such evidence as is available, adoption or adaptation is to be entertained; it might be wondered whether it is useful to apparently preclude learner

production or simply not to force it; the extrapolator may wish to think how best to anticipate teachers' anxieties, to speculate about the use of groupwork, or the likely effects of a greater rhetorical range allowed for in task selection. An extrapolator might simply ask the question: if I were to teach in the ways suggested by the CTP would it work with my students in my school.

The starting point for all such reasoning would be a judgment of the credibility of each of the findings presented in the present evaluation study. Thus, this study does not expect to contribute to theories of language learning, but attempts to provide information that may assist in pointing the way to future programs.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abt, W.P. and Magidson, J. 1980. Reforming schools: problems in program implementation and evaluation. Beverly Hills: Sage.

Agard, F.B. and Dunkel, H.B. 1948. An investigation of second language teaching. Boston: Ginn and Company.

Alkin, M.C. 1969. Evaluation theory development. Evaluation Comment 2 (1): 2-7.

Allen, J.P.B., Cummins, J., Mougeon, R., and Swain, M. 1983. The development of bilingual proficiency. Unpublished second year report, Ontario Institute for Studies in Education, Toronto.

Allport, G.W. 1968. The historical background of modern psychology. In G. Lindzey and E. Aronson (Eds.). The handbook of social psychology. Vol.1.. Reading, MA: Addison-Wesley.

Allwright, R.L. 1972. Prescription and description in the teaching of language teachers. In Proceedings of the third International Congress of Applied Linguistics (AILA, Copenhagen), 150-156. Heidelberg: Julius Groos Verlag.

Allwright, R.L. 1975. Problems in the study of the language teacher's treatment of learner error. In M.K. Burt and H.C. Dulay (Eds.). On TESOL '75: New directions in second language learning, teaching and bilingual education, 96-109. Washington, D.C.: TESOL.

Allwright, R.L. 1977. Turns, topics and tasks: patterns

of participation in language learning and language teaching. Paper presented at the 11th annual TESOL Convention, Miami, Florida, April.

Allwright, R.L. 1981. Report on visit to India. London: British Council.

Allwright, R.L. 1983. Classroom-centred research on language teaching and learning: a brief historical overview. TESOL Quarterly 17 (2): 191-204.

Anderson, L.M., Evertson, C.M., and Brophy, J.E. 1979. An experimental study of effective teaching in first grade reading groups. Elementary School Journal 79: 193-223.

Anderson, R.B., St. Pierre, R.G., Proper, E.C., and Stebbins, L.B. 1978. Pardon us, but what was the question again? A response to the critique of the Follow Through evaluation. Harvard Educational Review 48: 161-170.

Ankers, J. 1974. Commentary on paper by P.S. Green. In Role et efficacite du laboratoire de langues dans l'enseignement secondaire et universitaire. No. speciale, No. 20, Bulletin CILA.

Antony, E.M. 1963. Approach, method and technique. English Language Teaching 17: 63-67.

APA (American Psychological Association). 1954. Technical recommendations for psychological techniques and diagnostic techniques. Washington, D.C.: Author.

APA (American Psychological Association). 1966. Standards

for educational and psychological tests and manuals. Washington, D.C.: Author.

Argyle, M. 1969. Social Interaction. Chicago: Atherton Press.

Asher, J.J. 1966. The learning strategy of the Total Physical Response: a review. Modern Language Journal 50: 79-84.

Asher, J.J. 1972. Children's first language as a model for second language learning. Modern Language Journal 56 (3): 133-139.

Asher, J.J., Kusudo, J.A., and de la Torre, R. 1974. Learning a second language through commands: the second field test. Modern Language Journal 58: 24-32.

Atkin, M. 1968. Behavioral objectives in curriculum design: a cautionary note. The Science Teacher 35: 27-30.

Ausubel, D.P. 1964. Adults versus children in second language learning: psychological considerations. Modern Language Journal 48: 420-424.

Azrin, N.H. 1977. A strategy for applied research: learning based but outcome oriented. American Psychologist 32: 140-149.

Babbie, E.R. 1975. The practice of social research. Belmont, Calif.: Wadsworth.

Bailey, L.G. 1975. An observational method in the foreign language classroom: a closer look at interaction analysis. Foreign Language Annals 8 (4): 335-344.

Bales, R.F. 1951. Interaction process analysis: a method

for the study of small groups. Cambridge, MA: Addison-Wesley.

Ball, S. and Bogatz, G.A. 1970. The first year of Sesame Street: an evaluation. Princeton, NJ: Educational Testing Service.

Bannister, D. 1966. Psychology as an exercise in paradox. Bulletin of the British Psychological Society 19: 21-26.

Barker, R.G. 1965. Explorations in ecological psychology. American Psychologist 20: 1-14.

Barker, R.G., Dembo, T., and Lewin, K. 1941. Frustration and regression: a study of young children. University of Iowa Studies in Child Welfare 18, No.1.

Barkman, B. 1978. Classroom interaction. In P.M. Lightbown and B. Barkman Interactions among learners, teachers, texts, and methods of English as a second language. Progress Report 1977-1978, 71-78. Montreal: Concordia University.

Bartko, J.J. 1976. On various intraclass correlation reliability coefficients. Psychological Bulletin 83 (5): 762-765.

Bathory, Z. 1977. The field-trial stage of curriculum evaluation. In A.Lewy (Ed.). Handbook of Curriculum Evaluation. Paris: UNESCO / New York: Longman.

Bellack, A.A., Kliebard, H.M., Hyman, R.T., and Smith, F.L.Jnr. 1966. The language of the classroom. New York: Teachers College Press.

Berkowitz, L. and Donnerstein, E. 1982. External validity is more than skin-deep: some answers to criticisms of laboratory experiments. American Psychologist 37 (3): 245-257.

Berliner, D.C. 1975. Impediments to the study of teacher effectiveness. Paper presented at the National Invitational Conference on Research on Teacher Effects: an Examination by Decisionmakers and Researchers, Austin. (ERIC Document Reproduction Service No. ED 128 343).

Bialystok, E., Frohlich, M., and Howard, J. 1979. Studies on second language learning and teaching in classroom settings: strategies, processes, and functions. Unpublished report, Ontario Institute for Studies in Education, Toronto.

Billows, F.L. 1955. Suggestions for increasing the efficiency of English teaching in South India. Teaching English: a Magazine Devoted to the Teaching of English in South India 2 (3): 21-23.

Birkmaier, E.M. and Lange, D. 1967. Foreign language instruction. Review of Educational Research 37: 186-199.

Blackith, R.E. and Reyment, R.A. 1971. Multivariate morphometrics. London: Academic Press.

Borich, G.D. and Klinzing, G. 1984. Some assumptions in the observation of classroom process with suggestions for improving low inference measurement. Journal of Classroom Interaction 20 (1): 36-44.

Bose, M.N.K. 1980. A teacher's experience. In RIE Bulletin q.v., 4 (1): p.81-88.

Bracht, G.H. and Glass, G.V. 1968. The external validity of experiments. American Educational Research Journal 5 (4): 437-474.

Brownell, W.A. 1966. The evaluation of learning under dissimilar systems of instruction. California Journal of Educational Research 17: 80-90.

Brumfit, C.J. 1980. The role of research: some experimental investigations into language teaching methodology, and some of their limitations. Ch.7. in Problems and principles in English Teaching. Oxford: Pergamon Press.

Brumfit, C.J. 1984. The Bangalore procedural syllabus. English Language Teaching Journal 38 (4): 233-241.

Brunswik, E. 1955. Representative design and probabilistic theory in a functional psychology. Psychological Review 62 (3): 193-217.

Bryk, A.S. 1981. Disciplined inquiry or policy argument? Harvard Educational Review 51 (4): 497-509.

Burt, M., Dulay, H.C., and Hernandez, E. 1973. Bilingual Syntax Measure. New York: Harcourt Brace Jovanovich.

Bushman, R.W. and Madsen, H.S. 1976. A description and evaluation of suggestopedia - a new teaching methodology. In J.F. Fanselow and R.H. Crymes (Eds.). OnTESOL '76, 29-40. Washington, D.C.: TESOL.

Campbell, D.T. and Stanley, J.C. 1963. Experimental and

quasi-experimental designs for research on teaching. In N.L. Gage (Ed.). Handbook of research on teaching, 171-246. Chicago: Rand McNally.

Carlsmith, J.M., Ellsworth, P.C., and Aronson, E. 1976. Methods of research in social psychology. Reading, MA: Addison-Wesley.

Capelle, G.C., Jarvella, R.J., and Revelle, E. n.d. Development of computer-assisted observational system for teacher training. Center for Research on Language and Language Behavior, University of Michigan.

Carroll, J.B. 1963. Research on teaching foreign languages. In N.L Gage (Ed.). Handbook of research on teaching, 1060-1100. Chicago, Rand McNally.

Carroll, J.B. 1965. The contributions of psychological theory and educational research to the teaching of foreign languages. Modern Language Journal 49 (5): 273-281.

Carroll, J.B. 1969. What does the Pennsylvania foreign language research project tell us? Foreign Language Annals 3 (2): 214-236.

Carroll, J.B. and Sapon, S. 1959. The Modern Language Aptitude Test. New York: The Psychological Corporation.

Carton, A.S. 1966. The 'method of inference' in foreign language study. The Office of Research and Evaluation, The Division of Teacher Education, The City University of New York.

Casey, J.B. 1968. The effectiveness of two methods of

teaching English as a foreign language in some Finnish secondary schools. Unpublished report, University of Helsinki.

Cattell, R.B. 1966. Guest editorial: Multivariate behavior research and the integrative challenge. Multivariate Behavioral Research 1: 4-23.

Charters, W.W. and Jones, J. 1973. On the risks of appraising non-events in program evaluation. Educational Researcher 2 (11): 5-7.

Charters, W.W. and Pellegrin, R. 1973. Barriers to the innovation process: four case studies of differentiated staffing. Administrative Science Quarterly 9 (1): 3-14.

Chastain, K.D. 1970. A methodological study comparing the audio-lingual habit theory and the cognitive code-learning theory - a continuation. Modern Language Journal 54 (4): 257-266.

Chastain, K.D. and Woerdehoff, F.J. 1968. A methodological study comparing the audio-lingual habit theory and the cognitive code-learning theory. Modern Language Journal 52 (5): 268-279.

Chatfield, C. and Collins, A.J. 1980. Introduction to multivariate analysis. London: Chapman and Hall.

Chaudron, C. 1977. A descriptive model of discourse in the corrective treatment of learners' errors. Language Learning 27 (1): 29-46.

Chaudron, C. 1986. Teachers' priorities in correcting

learners' errors in French immersion classes. In R.R. Day (Ed.). Talking to learn: conversation in second language acquisition, 64-84. Rowley, MA: Newbury House.

Churchman, D. 1979. A new approach to evaluating the implementation of innovative educational programs. Educational Technology 19: 25-28.

Cicirelli, V.G. et al. 1969. The impact of Head Start: an evaluation of the effects of Head Start on children's cognitive and affective development. Study by Westinghouse Learning Corporation and Ohio University. Washington, D.C.: Office of Economic Opportunity.

Clark, J.L.D. 1969. The Pennsylvania project and the 'audiolingual versus traditional' question. Modern Language Journal 53 (6): 388-396.

Coladarci, T. and Gage, N.L. 1984. Effects of a minimal intervention on teacher behavior and student achievement. American Educational Research Journal 21 (3): 539-555.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. and York, R. 1966. Equality of educational opportunity. Washington, D.C.: U.S. Office of Health, Education, and Welfare.

Collingham, M. 1981. Teaching techniques in task-based learning. M.A. dissertation, Lancaster University.

Comber, L.C. and Keeves, J.P. 1973. Science education in 19 countries: an empirical study. International Studies in Evaluation. Stockholm: Almqvist and

Wiksell.

Cook, T.D. and Campbell, D.T. 1976. The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.). Handbook of industrial and organisational psychology. Chicago: Rand McNally.

Cook, T.D. and Campbell, D.T. 1979. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally.

Cook, T.D. and Reichardt, C.S. 1979. (Eds.). Qualitative and quantitative methods in evaluation research. Beverly Hills, Calif.: Sage.

Corder, S.P. 1968. Advanced study and the experienced teacher. In G.E. Perren (Ed.). Teachers of English as a second language: their training and preparation. Cambridge: Cambridge University Press.

Corder, S.P. 1984. Review of Krashen. Applied Linguistics 5 (1): 56-58.

Cornfield, J. and Tukey, J.W. 1956. Average values of mean squares in factorials. Annals of Mathematical Statistics 27: 907-949.

Courchene, R. 1980. The error analysis hypothesis, the contrastive analysis hypothesis, and the correction of error in the second language classroom. TESL Talk 11 (2): 3-13 / 11 (3): 10-29.

Crandall, D.P., Bauchner, J.E., Loucks, S.F. and Schmidt, W.H. 1982. Models of the school improvement process:

factors contributing to success. A study of dissemination efforts supporting high school improvement. Paper presented at the annual meeting of the American Educational Research Association, New York, March. (ERIC Document Reproduction Service No. ED 251 918).

Crawford, J., Gage, N.L., Corno, L., Stayrook, N., and Mitman, A. 1978. An experiment on teacher effectiveness and parent-assisted instruction in the third grade, vols. 1-3. Stanford, Calif.: Program on Teaching Effectiveness, Center for Educational Research, Stanford University.

Cronbach, L.J. 1963. Course improvement through evaluation. Teachers College Record 64: 672-683.

Cronbach, L.J. 1975. Beyond the two disciplines of scientific psychology. American Psychologist 30 (2): 116-127.

Cronbach, L.J. 1982. Designing evaluations of educational and social programs. San Francisco: Jossey Bass.

Cronbach, L.J. 1984. Essentials of psychological testing. Fourth Edition. New York: Harper and Row.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., and Weiner, S.S. 1980. Toward reform of program evaluation. San Francisco: Jossey-Bass.

Cronbach, L.J., Rajaratnam, M., and Gleser, G. 1963. Theory of generalisability: a liberation of reliability theory. British Journal of Statistical Psychology

16: 137-163.

Cronbach, L.J. and Suppes, P. 1969. Research for tomorrow's schools: disciplined inquiry in education. New York: Macmillan.

Cronbach, L.J. and Snow, R.E. 1981. Aptitudes and instructional measures (2nd Edition). New York: Irvington.

Crookes, G. 1986. Task classification: a cross-disciplinary review. Technical Report No.4., University of Hawaii at Manoa.

Crowther, F. 1972. Factors affecting the rate of adoption of the 1971 Alberta social studies curriculum for elementary schools. M.A. thesis, University of Alberta.

Daillak, R.H. 1983. Evaluators in an urban school district: organisational constraints upon evaluation influence. Studies in Educational Evaluation 9: 33-46.

Davies, A. 1977. Introduction. In J.P.B. Allen and A. Davies (Eds.). The Edinburgh course in applied linguistics. Vol.4. Testing and experimental methods, 1-10. Oxford: Oxford University Press.

Davies, A. 1983. Evaluation and the Bangalore/Madras Communicational Teaching Project. Unpublished manuscript, University of Edinburgh, Department of Applied Linguistics.

Davies, A. 1984. Review of Oller: Issues in language testing research. Language Testing 1 (1): 111-115.

Delamont, S. 1976. Interaction in the classroom. London: Methuen.

Diaz-Guerrero, R., Reyes-Lagunes, I., Witzke, D., and Holzman, W.H. 1976. Plaza Sesamo in Mexico. Journal of Communication 26: 145-153.

Doerfler, L.A. and Chaplin, W.F. 1985. Type III error in research on interpersonal models of depression. Journal of Abnormal Psychology 94 (2): 227-230.

Downey, L. and Associates. 1975. The social studies in Alberta - 1975. Edmonton, Alberta: L. Downey Research Associates.

Dunkel, H.B. 1948. Second Language Learning. Boston: Ginn and Company.

Ebel, R.L. 1951. Estimation of the reliability of ratings. Psychometrika 16: 407-424.

Educational Testing Service. 1939-1941. Cooperative tests of German and French. Princeton, NJ: Author.

Eisner, E.W. 1977. On the uses of educational connoisseurship and criticism for evaluating classroom life. Teachers College Record 78 (3): 345-358.

Eisner, E.W. 1979. The educational imagination: on the design and evaluation of school programs. New York: Macmillan.

Eisner, E.W. 1984. Can educational research inform educational practice? Phi Delta Kappan 65 (7): 447-452.

Elashoff, J.D. 1969. Analysis of covariance: a delicate instrument. American Educational Research Journal 6

(3): 383-401.

Elia, T. 1981. Communicational teaching and the structuralist teacher. In RIE Bulletin 5 (1): 27-36.

Essef, P. 1970. Reported comments in P.D. Smith, Jnr. q.v.

Fanselow, J.F. 1977a. Beyond RASHOMON - conceptualising and describing the teaching act. TESOL Quarterly 11 (1): 17-39.

Fanselow, J.F. 1977b. The treatment of error in oral work. Foreign Language Annals 10 (5): 583-593.

Fawl, C.L. 1963. Disturbances experienced by children in their natural habitats. In R.G. Barker (Ed.). The stream of behavior, 99-126. New York: Appleton-Century-Crofts.

Fenstermacher, G.D. 1979. A philosophical consideration of recent research on teacher effectiveness. Review of Research in Education 6: 157-185.

Fenstermacher, G.D. 1982. On learning to teach effectively from research on teacher effectiveness. Journal of Classroom Interaction 17 (2): 7-12.

Fink, S.R. 1972. Dialog memorisation in introductory language instruction. A comparison of three different strategies. In Proceedings of the third International Congress of Applied Linguistics (AILA, Copenhagen), 273-289. Heidelberg: Julius Groos Verlag.

Fisher, R.A. 1966. The design of experiments. (8th Edition). Edinburgh: Oliver and Boyd.

Flanders, N.A. 1960. Interaction analysis in the classroom: a manual for observers. University of Michigan, Ann Arbor.

Forrester, D.L. 1975. Other research into the effectiveness of language laboratories. In P.S. Green (Ed.). q.v., 5-33.

Forrester, J. 1954. The syllabus in English for the Madras state. Teaching English: a Magazine Devoted to the Teaching of the English Language in India 1 (2): 35-45.

Forrester, J. 1955. Group work in English. Teaching English: a Magazine Devoted to the Teaching of English in India 2 (3): 7-10.

Fraser, B.J. 1984. Directions in curriculum evaluation. Studies in Educational Evaluation 10: 125-134.

Freedman, E.S. 1971. The road from Pennsylvania - where next in foreign language exploration? Audio-Visual Language Journal 9 (1): 33-38.

Freedman, E.S. 1976. Experimentation into foreign language teaching methodology. System 4 (1): 12-28.

Freedman, E.S. 1979. Valid research into foreign language teaching: two recent projects. System 7 (3): 187-199.

Freedman, E.S. 1982. Experimentation into foreign language teaching methodology. The research findings. System 10 (2): 119-133.

Freeman, H.E. 1964. Conceptual approaches to assessing impacts of large-scale intervention programs. Proceedings of the American Statistical Association (Social

Statistics Section), 192-198.

Freudenstein, R. 1976. How to analyse a foreign language lesson. Paper presented at the 10th annual TESOL Convention, New York, March.

Frick, T. and Semmel, M.I. 1978. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research 48 (1): 157-184.

Frohlich, M., Spada, N., and Allen, P. 1985. Differences in the communicative orientation of L2 classrooms. TESOL Quarterly 19 (1): 27-57.

Fullan, M. 1983. Evaluating program implementation: what can be learned from Follow Through. Curriculum Inquiry 13 (2): 215-227.

Fullan, M. and Pomfret, A. 1977. Research on curriculum and instruction implementation. Review of Educational Research 47 (1): 335-397.

Fuller, F.F. 1969. Concerns of teachers: a developmental conceptualisation. American Educational Research Journal 6 (2): 207-226.

Gabriel, J. 1957. An analysis of the emotional problems of the teacher in the classroom. Melbourne: F.W. Cheshire.

Gadlin, H. and Ingle, G. 1975. Through the one-way mirror: the limits of experimental self-reflection. American Psychologist 30: 1003-1009.

Gage, N.L. 1978. The scientific basis of the art of teaching. New York: Teachers College Press.

Garrison, J.W. and Macmillan, C.J.B. 1984. A philosophical critique of process-product research on teaching. Educational Theory 34 (3): 255-274.

Gary, J.O. 1975. Delayed oral practice in initial stages of second language learning. In M. Burt and H. Dulay (Eds.). On TESOL '75, 89-95. Washington, D.C.: TESOL.

Genesee, F. 1983. Bilingual education of majority language children: the immersion experiments in review. Applied Psycholingistics 4: 1-46.

Gibbons, J. 1985. The silent period: an examination. Language Learning 35: 255-267.

Gilpin, A. 1981. The influence of the teacher's method on the pre-task phase of problem-solving tasks in L1 and L2 classrooms. M.A. dissertation, Lancaster University.

Good, T. 1979. Teacher effectiveness in the elementary school. Journal of Teacher Education 30: 52-64.

Good, T. and Power, C. 1976. Designing successful classroom environments for different types of students. Journal of Curriculum Studies 8: 45-60.

Good, T. and Grouws, D.A. 1979. The Missouri mathematics effectiveness project: an experimental study in fourth grade classrooms. Journal of Educational Psychology 71 (3): 355-362.

Green, D.R. 1983. Content validity of standardised achievement tests and test curriculum overlap. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, March.

(ERIC Document Reproduction Service No. ED 235 237).

Green, P.S. 1972. A study of the effectiveness of the language laboratory in school, conducted at Archbishop Holgate's Grammar School, York, 1967-1970. IRAL 10 (3): 283-292.

Green, P.S. (Ed.). 1975. The language laboratory in school: performance and prediction. An account of the York study. Edinburgh: Oliver and Boyd.

Green, P.S. and Hawkins, E.W. 1975. Conclusions. In P.S. Green (Ed.). q.v.

Greenwood, J. 1985. Bangalore revisited: a reluctant complaint. English Language Teaching Journal 39 (4): 268-273.

Gross, N., Giacquinta, J., and Bernstein, M. 1971. Implementing organisational innovations: a sociological analysis of planned educational change. New York: Basic Books.

Guba, E.G. 1981. The paradigm revolution in inquiry: implications for vocational research and development. Paper presented at the National Center for Research in Vocational Education Staff Development Seminar, Columbus, Ohio. (ERIC Document Reproduction Service No. ED 212 829).

Hall, G.E. and Loucks, S.F. 1977. A developmental model for determining whether the treatment is actually implemented. American Educational Research Journal 14 (3): 263-276.

Hall, G.E. and Rutherford, W.W. 1983. Client concerns: a guide to facilitating institutional change. Austin, Texas: Research and Development Center for Teacher Education, University of Texas. (ERIC Document Reproduction Service No. ED 251 728).

Hall, G.E. and George, A. 1979. Stages of concern about the innovation: the concept, initial verification, and some implications. 1st draft. Austin, Texas: Research and Development Center for Teacher Education, University of Texas. (ERIC Document Reproduction Service No. ED 187 716).

Hall, G.E., George, A., and Rutherford, W.W. 1977. Measuring stages of concern about the innovation: a manual for use of the SoC questionnaire. Austin, Texas: Research and Development Center for Teacher Education, University of Texas. (ERIC Document Reproduction Service No. ED 147 342).

Hammerly, H.M. 1982. Synthesis in second language teaching: an introduction to languistics. Blaine, Washington: Second Language Publications.

Hanson, R.A. and Bailey, J.D. 1983. Program-fair educational evaluation and empirical curriculum inquiry. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development. (ERIC Document Reproduction Service No. ED 251 505).

Hanson, R.A., Schutz, R.E., and Bailey, J.D. 1977. Program-fair evaluation of instructional programs: initial results of the kindergarten reading readiness

inquiry. Los Alamitos, CA: Southwest Regional Laboratory for Educational Research and Development. (ERIC Document Reproduction Service No. ED 161 519).

Harre, R. and Secord, P.F. 1972. The explanation of social behavior. Oxford: Blackwell.

Hatch, E. and Farhady, H. 1982. Research design and statistics for applied linguistics. Rowley, MA: Newbury House.

Hauptman, P.C. 1971. A structural approach versus a situational approach to foreign language teaching. Language Learning 21 (2): 234-244.

Hawkins, E. 1975. The design of the York study. In P.S. Green (Ed.). q.v., 34-53.

Hawkins, L.E. 1971. Immediate versus delayed presentation of foreign language script. Modern Language Journal 55 (5): 280-290.

Haynes, S.N. and Wilson, C.C. 1979. Behavioral assessment: recent advances in methods, concepts, and applications. San Francisco: Jossey-Bass.

Henderson, T.H., Gomes, P.I., and Patton, M.Q. 1983. User-focused evaluation: a Caribbean example. Studies in Educational Evaluation 9: 47-58.

Henning, G. 1986. Quantitative methods in language acquisition research. TESOL Quarterly 20 (4): 701-708.

Henning, G. 1987. A guide to language testing: development, evaluation, research. Cambridge: Newbury House.

Hensman, B. 1955. Language teaching through group

activities. Teaching English: a Magazine Devoted to the Teaching of English in India 2 (3): 21-23.

Herbert, J and Attridge, C. 1975. A guide for developers and users of observation systems and manuals. American Educational Research Journal 12 (1): 1-20.

Hess, R. and Buckholdt, D. 1974. Degree of implementation as a critical variable in program evaluation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April.

Hilgard, E.R. and Bower, G.H. 1966. Theories of learning. (3rd Edition. New York: Appleton-Century-Crofts.

Hills, M. 1977. Book review. Applied Statistics 26: 339-340.

Hilton, M. 1969. A scientific experiment? Audio-Visual Language Journal 7 (2): 97-101.

Hocking, E. 1969. The laboratory in perspective: teachers, strategies, outcomes. Modern Language Journal 53 (6): 404-410.

House, E.R., Glass, G.V., Maclean, L.D., and Walker, D.F. 1978. No simple answer: critique of the Follow Through evaluation. Harvard Educational Review 48 (2): 128-160.

House, E.R., Thurston, P., and Hand, J. 1984. Evaluation reflections: the adversary hearing as a public forum. Studies in Educational Evaluation 10: 111-123.

Howatt, A.P.R. 1984. A history of English language teaching Oxford: Oxford University Press.

Jarvis, G.A. 1968. A behavioral observation system for

classroom foreign language skill acquisition activities. Modern Language Journal 52 (6): 335-341.

Jarvis, G.A. and Adams, S.J. 1979. Evaluating a second language program No.19 in Language in Education: Theory and Practice series. Arlington, Virginia: Center for Applied Linguistics.

Jenkins, D. 1976. Curriculum evaluation. Milton Keynes: The Open University.

Johnson, K. 1982. Communicative syllabus design and methodology. Oxford: Pergamon.

Johnson, R.K., Chan, R.M.K., and Shek, C.K.W. 1985. Language across the curriculum in teacher education. Paper presented at the 20th RELC Conference, Singapore, April.

Johnston, O.W. 1980. Implementing the intensive language model: an experiment in German at the University of Florida. Foreign Language Annals 13 (2): 99-106.

Joint Committee on Standards for Educational Evaluation. 1981. Standards for evaluations of educational programs, projects and materials. New York: McGraw-Hill.

Kasper, G. 1986. Repair in foreign language teaching. In Learning, teaching and communication in the foreign language classroom, 23-41. Aarhus: Aarhus University Press.

Keating, R.F. 1963. A study of the effectiveness of language laboratories. New York: Institute of Administrative Research, Teachers College.

Kempthorne, O. 1961. The design and analysis of experiments with some reference to educational research. In R.O. Collier, Jnr., and S.M. Elam (Eds.). Research design and analysis: second annual Phi Delta Kappa Symposium on Educational Research, 97-126. Bloomington, Indiana: Phi Delta Kappa.

Kerlinger, F.N. 1977. The influence of research on education practice. Educational Researcher 6: 5-12.

King, J.A. 1982. Studying the local use of evaluation: a discussion of theoretical issues and an empirical study. Studies in Educational Evaluation 8: 175-183.

Koehler, V. 1978. Classroom process research: present and future. Journal of Classroom Interaction 13 (2): 3-11.

Kounin, J. 1970. Discipline and group management in classrooms. New York: Holt, Rinehart and Winston.

Krashen, S.D. 1981. Second language acquisition and second language learning. Oxford: Pergamon Press.

Krashen, S.D. 1982. Principles and practice in second language acquisition. Oxford: Pergamon Press.

Krashen, S.D. and Terrell, T.D. 1983. The Natural Approach: language acquisition in the classroom. Oxford: Pergamon Press.

Kratchowill, T.R. 1979. Intensive research: a review of methodological issues in clinical, counseling, and school psychology. Review of Research in Education 7: 46-91.

Krumm, H.-J. 1973. Interaction analysis and microteaching for the training of foreign language teachers. IRAL 11

(2): 163-170.

Kumaruvadivelu, B. 1981. Turn organisation in L1 and L2 classroom discourse. M.A. dissertation, Lancaster University.

Lado, R. 1970. Reported comments In P.D. Smith, Jnr. 1970. q.v.

Leinhardt, G. 1980. Modeling and measuring educational treatment in evaluation. Review of Educational Research 50 (3): 393-420.

Leithwood, K.A. and Montgomery, D.J. 1980. Evaluating program implementation. Evaluation Review 4: 193-214.

Levin, L. 1972. Comparative studies in foreign language teaching: the GUME project. Stockholm: Almqvist and Wiksell.

Leviton, L.C. and Hughes, E.F.X. 1981. Research on the utilisation of evaluations: a review and synthesis. Evaluation Review 5: 525-548.

Lim, K.-B. 1968. Prompting versus confirmation, pictures versus translations, and other variables in children's learning of grammar in a second language. Unpublished thesis, Harvard University.

Lincoln, Y.S. and Guba, E.E. 1985. Research, evaluation, and policy analysis: heuristics for disciplined inquiry. (ERIC Document Reproduction Service No. ED 252 966).

Lindquist, E.F. 1953. Design and analysis of experiments in psychology and education. Boston: Houghton-Miflin.

Long, M.H. 1980. Inside the 'black box': methodological issues in classroom research on language learning. Language Learning 30 (1): 1-42. Reprinted in H.W. Seliger and M.H. Long (Eds.). 1983. Classroom oriented research in second language acquisition, 3-35. Rowley, MA: Newbury House.

Long, M.H. 1984. Process and product in ESL program evaluation. TESOL Quarterly 18 (3): 409-425.

Long, M.H. 1985. A role for instruction in second language acquisition: task-based language teaching. In K. Hyltenstam and M. Pienemann (Eds.). Modelling and assessing second language acquisition. London: Multilingual Matters.

Long, M.H., Adams, L., McLean, M., and Castanos, F. 1976. Doing things with words - verbal interaction in lockstep and small group classroom situations. In J.F. Fanselow and R.H. Crymes (Eds.). On TESOL '76, 137-153. Washington, D.C.: TESOL.

Loucks, S.F. 1983. Defining fidelity: a cross-study analysis. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April. (ERIC Document Reproduction Service No. ED 249 659).

Loucks, S.F., Newlove, B.N, and Hall, G.E. 1975. Measuring levels of use of the innovation: a manual for trainers, interviewers and raters. Austin, Texas: Research and Development Center for Teacher Education, University of Texas.

Loucks, S.F. and Melle, M. 1981. The implementation of a district-wide science curriculum: the effects of a three year effort. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April. (ERIC Document Reproduction Service No. ED 204 181).

Lukas, C. and Wohleb, C. 1973. Implementation of Head Start planned variation: 1970-71, Part 1 and 2. Cambridge, MA., Huron Institute.

Mahoney, M.J. 1978. Experimental methods and outcome evaluation. Journal of Consulting and Clinical Psychology 46: 660-672.

Marston, D., Deno, S., and Tindal, G. 1983. A comparison of standardised achievement tests and direct measurement techniques in measuring pupil progress. Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED 236 198).

Martin, D.S. and Saif, P.J. 1984. Curriculum change from the grass-roots. Paper presented at the annual meeting of the Association for Supervision and Curriculum Development, Chicago, March. (ERIC Document Reproduction Service No. ED 254 913).

McArdle, R.J. and Scebold, C.E. 1968. Report on a workshop seminar held at the University of Nebraska, Summer 1968. Modern Language Journal 52 (8): 502-503.

McEwan, N.Z. 1976. An exploratory study of the multi-

dimensional nature of teacher-student verbal interaction in second language classrooms. Unpublished thesis, University of Alberta.

McFarlane, J.M. 1975. Focus analysis. (Some notes towards a system of FL classroom observation). In R.L. Allwright (Ed.). Working papers: language teaching classroom research, 131-145. University of Essex, Department of Language and Linguistics.

McGaw, B., Wardrop, J.L., and Bunda, M.A. 1972. Classroom observation schemes: where are the errors? American Educational Research Journal 9 (1): 13-27.

McIntyre, D. and Mitchell, R. 1983. Processes, outcomes and context of Bilingual education in the Western Isles. University of Stirling, Department of Education.

McKinnon, K.R. 1965. An experimental study of the learning of syntax in second language learning. Unpublished thesis, Harvard University.

McLaughlin, M.W. and Berman, P. 1975. Macro and micro implementation. The Rand Corporation, Santa Monica, CA. (ERIC Document Reproduction Service No. ED 118 477).

Medley, D.M. 1977. Future directions for process-product research. Journal of Classroom Interaction 13 (1): 3-8.

Medley, D.M. 1978. Research in teacher effectiveness - where it is and how it got there. Journal of Classroom Interaction 13 (2): 16-21.

Medley, D.M. 1982a. Systematic observation. In H.E. Mitzel (Ed.). Encyclopedia of educational research. 5th Edition, 1841-1851. New York: The Free Press.

Medley, D.M. 1982b. Teacher effectiveness. In H.E. Mitzel (Ed.). Encyclopedia of educational research. 5th Edition, 1894-1903. New York: The Free Press.

Medley, D.M. and Mitzel, H.E. 1958. A technique for measuring classroom behavior. Journal of Educational Psychology 49: 86-92.

Medley, D.M. and Mitzel, H.E. 1963. Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.). Handbook of research on teaching, 247-328. Chicago: Rand McNally.

Medley, D.M., Soar, R., and Coker, H. 1983. The minimum conditions for valid evaluation of teacher performance. Journal of Classroom Interaction 19 (1): 22-27.

Mehan, H. 1979. Learning lessons: social organisation in the classroom. Harvard: Harvard University Press.

Mitchell, R. 1977. A review of systematic observation instruments used for the analysis of FL classroom interaction. Unpublished paper, Department of Education, University of Stirling.

Mitchell, R., Parkinson, B., and Johnstone, R. 1981. The foreign language classroom: an observation study. Stirling Educational Monographs No.9., Department of Education, University of Stirling.

Mitroff, I.I. and Blankenship, V. 1973. On the

methodology of the holistic experiment: an approach to the conceptualisation of large-scale social experiments. Journal of Technology Forecasting Social Change 4: 339-353.

Mitroff, I.I. and Featheringham, T.R. 1974. On systematic problem solving and the error of the third kind. Behavioral Science 19: 383-393.

Mizon, S. 1981. Teacher-talk. M.A. dissertation, Lancaster University.

Modern Language Association. 1963. MLA cooperative foreign language tests. Princeton, NJ: Educational Testing Service.

Mook, D.J. 1983. In defense of external invalidity. American Psychologist 38 (4): 379-387.

Moskowitz, G. 1971. Interaction analysis - a new modern language for supervisors. Foreign Language Annals 5 (2): 211-221.

Moskowitz, G. 1976. The classroom interaction of outstanding foreign language teachers. Foreign Language Annals 9: 135-143.

Mueller, T.H. 1971. The effectiveness of two learning models: the audiolingual habit theory and the cognitive code-learning theory. In P. Pimsleur and T. Quinn (Eds.). The psychology of second language learning, 113-122. Cambridge: Cambridge University Press.

Naiman, N., Frohlich, M., Stern, H.H., and Todesco, A. 1978. The good language learner. Toronto: Ontario Institute for Studies in Education.

Naumann-Etienne, M. 1974. Bringing about open education: strategies for innovation. Unpublished thesis, University of Michigan.

Nearhoof, O. 1969. Teacher-pupil interaction in the foreign language classroom: a technique for self-evaluation. Unpublished paper, quoted in F.M. Grittner (Ed.). Teaching foreign languages, 328-330. New York: Harper and Row.

Nevo, D. 1983. The conceptualisation of educational evaluation: an analytical review of the literature. Review of Educational Research 53 (1): 117-128.

Nevo, D. 1985. Face validity revisited. Journal of Educational Measurement 22 (4): 287-293.

Nystrom, N.Y. 1983. Teacher-student interaction in bilingual classrooms: four approaches to error feedback. in H.W. Seliger and M.H. Long (Eds.). Classroom oriented research in second language acquisition, 169-188. Rowley, MA: Newbury House.

O'Keefe, M. 1984. The impact of evaluation on federal education program policies. Studies in Educational Evaluation 10: 61-74.

Oller, J.W. 1979. Language tests at school. London: Longman.

Olsson, M. 1973. Learning grammar: an experiment. English Language Teaching 27 (3): 266-269.

Orne, M.T. 1962. On the social psychology of the psychological experiment: with particular reference to

demand characteristics and their implications. American Psychologist 17: 776-783.

Otto, F. 1969. The teacher in the Pennsylvania project. Modern Language Journal 53 (6): 411-420.

Owens, T.R. 1973. Educational evaluation by adversary proceedings. In E.R. House (Ed.). School evaluation: the politics and the process, 295-305. Berkeley, CA: McCutchan.

Pal, A. 1982. An applied psycholinguistic experiment in remedial teaching of English grammar. IRAL 20 (2): 152-160.

Palmer, H.E. 1921/1964. The principles of language study. London: Harrap. (Republished by Oxford University Press, 1964, edited by R. Mackin).

Parlett, M. and Hamilton, D. 1978. Evaluation as illumination: a new approach to the study of innovatory programmes. In D. Hamilton, D. Jenkins, C. King, B. MacDonald and M. Parlett (Eds.). Beyond the numbers game. London: Macmillan.

Patel, M.S. 1962. The structural syllabus at work in India. English Language Teaching 16 (3): 145-151.

Patton, M.Q. 1978. Utilization-focused evaluation. Beverly Hills, CA: Sage.

Patton, M.Q. 1982. Practical Evaluation. Beverly Hills: Sage.

Pereboom, A.C. 1971. Some fundamental problems in experimental psychology: an overview. Psychological Reports 28: 439-455.

Peterson, D., Micceri, T., and Smith, B.O. 1985. Measurement of teacher performance: a study in instrument development. Teaching and Teacher Education 1 (1): 63-77.

Phillips, D.C. 1981. Toward an evaluation of the experiment in educational contexts. Educational Researcher 10 (6): 13-20.

Phillips, M. 1932. Some problems of adjustment in the early years of a teacher's life. British Journal of Educational Psychology 2: 237-256.

Politzer, R.L. 1968. The role and place of the explanation in the pattern drill. IRAL 6 (4): 315-331.

Politzer, R.L. 1977. Foreign language teaching and bilingual education: implications of some recent research findings. Paper presented at the ACTFL Annual Conference, San Francisco, CA, November.

Postovsky, V.A. 1974. Effects of delay in oral practice at the beginning of second language learning. Modern Language Journal 58: 229-239.

Prabhu, N.S. 1980a. Theoretical background to the Bangalore project. In RIE Bulletin 4 (1), q.v., 17-26.

Prabhu, N.S. 1980b. Reactions and predictions. In RIE Bulletin 4 (1), q.v., 160-164.

Prabhu, N.S. 1980c. Methodological foundations of the Bangalore project. In RIE Bulletin 4 (1), q.v., 27-39.

Prabhu, N.S. 1981. Communicational Teaching Project: theoretical background. In RIE Bulletin 5 (1), q.v.,

1-8.

Prabhu, N.S. 1982.The Communicational Teaching Project, South India. Mimeo, British Council, Madras.

Prabhu, N.S. 1984. Coping with the unknown in language pedagogy. Paper presented at the British Council 50th Anniversary Seminar.

Prabhu, N.S. 1985. Correspondence (Response to Greenwood). English Language Teaching Journal 40 (1): 77.

Prabhu, N.S. 1987. Second Language Pedagogy. Oxford: Oxford University Press.

Prokop, M. 1974. Sequential analysis of FL verbal interaction. Alberta Journal of Educational Research 20 (4): 334-341.

Provus, M. 1971. Discrepancy evaluation. Berkeley, CA: McCutchan.

Raiffa, H. 1968. Decision analysis: introductory lectures on choices under uncertainty. Reading, MA: Addison-Wesley.

Rassias, J.A. 1971. New dimensions in language training: the Dartmouth College experiment. ADFL 3: 23-27.

Rice, E.G. and Higgins, N. 1982. The role of standardised testing in the teacher's evaluation strategy. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, March. (ERIC Document Reproduction Service No. ED 222 498).

Richards, J.C. 1984. The secret life of methods. TESOL Quarterly 18: 7-23.

Richards, J.C. and Rodgers, T. 1982. Method: approach,

design, and procedure. TESOL Quarterly 16 (2): 153-168.

Richards, J.C. and Rodgers, T. 1986. Approches and methods in language teaching: a description and analysis. New York: Cambridge University Press.

RIE (Regional Institute of English, South India). 1979. Newsletter. Vol.1. No.1. Bangalore: Author.

RIE (Regional Institute of English, South India). 1979. Newsletter. Vol.1. No.2. Bangalore: Author.

RIE (Regional Institute of English, South India). 1980. Newsletter. Vol.1. No.4. Bangalore: Author.

RIE (Regional Institute of English, South India). 1980. Bulletin. Vol.4. No.1. Bangalore: Author.

RIE (Regional Institute of English, South India). 1981. Bulletin. Vol.5. No.1. Bangalore: Author.

Riley, P. 1977. Discourse networks in classroom interaction: some problems in communicative language teaching. Paper presented at the BAAL seminar, University of Bath, April.

Roberts, J.T. 1982. Recent developments in ELT - Part II. Language Teaching: the International Abstracting Journal for Language Teachers and Applied Linguists 15 (3): 174-194.

Robson, C. 1973. Experiment, design and statistics in psychology. Middlesex: Penguin.

Rockwell, S.K. 1982. Evaluation abstract. A responsive approach to evaluating an alcohol program for youth.

Studies in Educational Evaluation 8: 175-183.

Rosenshine, B.V. and Furst, N. 1973. The use of direct observation to study teaching. In R.M.W. Travers (Ed.). Second handbook of research on teaching, 122-183. Chicago: Rand McNally.

Rosenthal, D. and Frank, J.D. 1956. Psychotherapy and the placebo effect. Psychological Bulletin 53: 294-302.

Ross, S. Forthcoming. Praxis and product in the EFL classroom. In A. Beretta and C.J. Alderson (Eds.). Evaluating language education programs. Cambridge: Cambridge University Press.

Rothfarb, S.H. 1970. Teacher-pupil interaction in the FLES class. Hispania 53 (2): 256-260.

Rowley, G.L. 1976. The reliability of observational measures. American Educational Research Journal 13 (1): 51-59.

Rutman, L and Mowbray, G. 1983. Understanding program evaluation. Beverly Hills: Sage.

Salica, C. 1981. Testing a model of corrective discourse. M.A. dissertation, University of California at Los Angeles.

Saraswathi, V. 1984. The Communicational Teaching Project: an evaluation. M.Sc. dissertation, Department of Applied Linguistics, University of Edinburgh.

Savignon, S.J. 1972. Communicative competence: an experiment in foreign language teaching. Philadelphia: The Center for Curriculum Development, Inc.

Schegloff, E.A., Jefferson, G., and Sacks, H. (1977). The

preference for self-correction in the organisation of repair in conversation. Language 53: 361-382.

Scherer, G.A.C. and Wertheimer, M. 1964. A psycholinguistic experiment in foreign language teaching. New York: McGraw-Hill.

Scott, W.A. 1955. Reliability of content analysis. Public Opinion Quarterly 19: 321-325.

Scriven, M.S. 1967. The methodology of evaluation. In R.Tyler, R. Gagne and M.S. Scriven Perspectives of curriculum evaluation. AERA Monograph Series on Curriculum Evaluation, no.1. Chicago: Rand McNally.

Scriven, M.S. 1972. Prose and cons about goal-free evaluation. Evaluation Comment 3. Los Angeles Center for the Study of Evaluation, University of California.

Scriven, M.S. 1981. Evaluation Thesaurus. 3rd Edition. Inverness, CA: Edgepress.

Secolsky, C. 1987. On the direct measurement of face validity: a comment on Nevo. Journal of Educational Measurement 24 (1): 82-83.

Seliger, H.W. 1975. Inductive method and deductive method in language teaching: a re-examination. IRAL 13 (1): 1-18.

Seliger, H.W. 1977. Does practice make perfect? A study of interaction patterns and L2 competence. Language Learning 27 (2): 263-278.

Sells, S.B. 1966. Ecology and the science of psychology. Multivariate Behavioral Research 1: 131-144.

Shaver, J.P. 1964. The ability of teachers to conform to two styles of teaching. The Journal of Experimental Education 32: 259-267.

Shaver, J.P. 1983. The verification of independent variables in teaching methods research. Educational Researcher 12 (8): 3-9.

Shoemaker, D.M. 1972. Evaluating the effectiveness of competing instructional programs. Educational Researcher 1: 5-8.

Siegel, S. 1956. Nonparametric statistics for the behavioral sciences. Tokyo: McGraw-Hill Kogakusha.

Simon, A. and Boyer, E.G. (Eds.). 1970. Mirrors for behavior: an anthology of classroom observation instruments. Supplementary volunes A and B. Philadelphia: Research for Better Schools.

Simon, A. and Boyer, E.G. 1974. Mirrors for behavior III: an anthology of classroom observation instruments. Wyncote, Pennsylvania: Communication Materials Center.

Sjoberg, K. and Trope, B. 1968. The value of external direction and individual discovery in learning situations: the learning of a grammatical rule. Scandinavian Journal of Educational Research 13: 233-240.

Smith, D.A. 1962. The Madras 'snowball': an attempt to retrain 27,000 teachers of English to beginners. English Language Teaching 17 (1): 3-9.

Smith, D.A. 1968. In-service training for teachers of English in developing countries. In G.E. Perren (Ed.). Teachers of English as a second language: their

training and preparation. Cambridge: Cambridge University Press.

Smith, E.R. and Tyler, R.W. 1942. Appraising and recording educational progress. New York: Harper and Row.

Smith, P.D. 1970. A comparison of the cognitive and audio-lingual approaches to foreign language instruction: the Pennsylvania foreign language project. Philadelphia: The Center for Curriculum Development, Inc.

Snow, R.E. 1974. Representative and quasi-representative designs for research on teaching. Review of Educational Research 44 (3): 265-291.

Soar, R.S. and Soar, R.M. 1982. Measurement of classroom process. In B. Spodek (Ed.). Handbook of research in early childhood education, 592-617. New York: The Free Press.

Soar, R.S., Soar, R.M., and Ragosta, M. 1971. Florida Climate and Control System. Gainesville: Institute for Development of Human Resources, University of Florida.

Stake, R.E. The countenance of educational evaluation. Teachers College Record 68: 523-540.

Stake, R.E. 1975. Evaluating the arts in education: a responsiveness approach. Columbus, Ohio: Merrill.

Stallings, J. 1975. Implications and child effects of teaching practices in Follow Through classrooms. Monographs of the Society for Research in Child Development, 40, No. 7-8, (Serial no. 163).

Stallings, J. 1977. How instructional processes relate to child outcomes. In G.D. Borich (Ed.). The appraisal of teaching: concepts and process. Reading, MA: Addison-Wesley.

Starr, W. 1970. Reported comments In P.D. Smith, Jnr. 1970. q.v., Appendix F.

Statistical Graphics Corporation. 1986. Statgraphics: statistical graphics system. User's guide. Rockville, Maryland: Author.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Anderson, R.B., and Cerva, T.R. 1977. Education as experimentation: a planned variation model, Vol.IV-A, an evaluation of Follow Through. Cambridge, MA: Abt Associates, Inc.

Stern, H.H. 1983. Fundamental concepts of language teaching. Oxford: Oxford University Press.

Stodolsky, S.S. 1984. Teacher evaluation: the limits of looking. Educational Researcher 13 (9): 11-18.

Stufflebeam, D.L. 1985. Coping with the point of entry problem in evaluating projects. Studies in Educational Evaluation 11: 123-129.

Stufflebeam, D.L., Foley, W.J., Gephart, W.J., Guba, E.G., Hammon, R.L., Merriman, H.O., amd Provus, M.M. 1971. Educational evaluation and decision-making. Ithaca, Illinois: Peacock.

Stufflebeam, D.L. and Webster, W.J. 1980. An analysis of alternative approaches to evaluation. Educational Evaluation and Policy Analysis 2 (3): 5-20.

Swain, M. and Lapkin, S. 1982. Evaluating bilingual education: a Canadian case study. Clevedon, Avon: Multilingual Matters.

Sweet, H. 1899/1964. The practical study of languages: a guide for teachers and learners. London: Oxford University Press.

Taba, H. 1966. Teaching strategies and cognitive functioning in elementary school children. Cooperative Research Project No. 2404, San Francisco State College.

Talmage, H. 1982. Evaluation of programs. In H.E. Mitzel (Ed.). Encyclopedia of educational research. 5th Edition, 592-611. New York: The Free Press.

Thiele, A. and Scheibner-Herzig, G. 1983. Listening comprehension training in teaching English to beginners. System 11 (3): 277-286.

Thouless, R.H. 1969. The teaching of foreign languages. In Map of educational research. Slough, Berks.: National Foundation for Educational Research in England and Wales.

Torrey, J.W. 1969. The learning of grammatical patterns. Journal of Verbal Learning and Verbal Behavior 8: 360-368.

Travers, R.M.W., Rabinowitz, W., and Nemovicher, E. 1952. The anxieties of a group of student teachers. Educational Administration and Supervision 38: 368-375.

Tucker, G.R., Lambert, W.E., and Rigault, A. 1969.

Students' acquisition of French gender distinctions: a pilot investigation. IRAL 7: 51-55.

Tyler, R.W. 1949. Basic principles of curriculum and instruction. Chicago: University of Chicago Press.

Ullmann, R. and Geva, E. 1982. The target language observation scheme (TALOS). York Region Board of Education, Core French Evaluation Project. Unpublished report, Ontario Institute for Studies in Education, Toronto.

Valdman, A. 1970. Evaluation of DPI/FL research projects. Appendix G in P.D. Smith. 1970. q.v., 359-363.

Valette, R.M. 1969. The Pennsylvania project, its conclusions and implications. Modern Language Journal 53 (6): 396-404.

Valette, R.M. 1970. Some conclusions to be drawn from the Pennsylvania study. Appendix H in P.D. Smith. 1970. q.v., 365-369.

Valette, R.M. 1971. Evaluation of learning in a second language. In B.S. Bloom, J.T. Hastings and G.F. Madaus (Eds.). Handbook of formative and summative evaluation in evaluation of student learning, 815-853.

Van Baalen, T. 1983. Giving learners rules: a study into the effect of grammatical instruction with varying degrees of explicitness. Interlanguage Studies Bulletin 7 (1): 71-100.

Von Elek, T. and Oskarsson, M. 1973. Teaching foreign language grammar to adults: a comparative study. Stockholm: Almqvist and Wiksell.

Wachtel, P.L. 1980. Investigation and its discontents: some constraints on progress in psychological research. American Psychologist 35: 399-408.

Wagner, M.J. and Tilney, G. 1983. The effect of 'superlearning techniques' on vocabulary acquisition and alpha brainwave production of language learners. TESOL Quarterly 17 (1): 5-17.

Walker, D.F. and Schaffarzick, J. 1974. Comparing curricula. Review of Educational Research 44 (1): 83-111.

Wang, M.C. and Ellett, C.D. 1982. Program validation: the state of the art. Topics in Early Childhood Special Education 1: 35-49.

Wang, M.C., Nojan, M., Strom, C.D., and Walberg, H.J. 1984. The utility of degree of implementation measures in program implementation and evaluation research. Curriculum Inquiry 14 (3): 249-286.

Weiss, C.H. 1972. Evaluation research: methods of assessing program effectiveness. Englewood Cliffs, NJ: Prentice-Hall.

Welch, W.W. 1983. Experimental inquiry and naturalistic inquiry: an evaluation. Journal of Research in Science Teaching 20 (2): 95-103.

Welch, W.W. and Walberg, H.J. A national experiment in curriculum evaluation. American Educational Research Journal 9 (3): 373-383.

Wesche, M.B. 1977. Learning behaviors of successful adult

students on intensive language training. Paper presented at the 1st Los Angeles Second Language Acquisition Research Forum, February.

Wholey, J. 1979. Evaluation: promise and performance. Washington, D.C.: Urban Institute.

Widdowson, H. 1978. Teaching language as communication. Oxford: Oxford University Press.

Wiley, D.E. 1969. A methodological review of the Pennsylvania foreign language research project. Foreign Language Annals 3 (2): 208-213.

Williams, D.L and Hull, W.L. 1968. Personal and situational variables which inhibit or stimulate the adoption of agricultural occupations curricula as an innovation in vocational agriculture by institute participants. Final Report. Stillwater: Oklahoma State University, Research Foundation.

Winitz, H. (Ed.). 1981. The comprehension approach to foreign language instruction. Rowley, MA: Newbury House.

Witkin, H.A., Moore, C.A., Goodenough, D.R. and Cox, P.W. 1977. Field-dependent and field-independent cognitive styles and their educational implications. Review of Educational Research 47 (1): 1-64.

Wohl, M. 1967. Classroom experiment to measure the relative efficiency of two different linguistic models in their application to the teaching of English as a foreign language. Unpublished report, Michigan University. (ERIC Document Reproduction Service No. ED

013 449).

Wolf, R.L. 1975. Trial by jury: a new evaluation method, 1: the process. Phi Delta Kappan 56: 185-187.

Wolf, R.M. (Guest Ed.). 1987. Educational evaluation: the state of the field. Introduction. International Journal of Educational Research 11 (1): 3-6.

Wolfe, D.E. and Jones, G. 1982. Integrating total physical response strategy in a level 1 Spanish class. Foreign Language Annals 15 (4): 273-280.

Wood, C.M. 1982. The role of standardised achievement tests in the management of instruction: a survey of teacher and administrator practices and attitudes. (ERIC Document Reproduction Service No. ED 219 446).

Woods, A., Fletcher, P., and Hughes, A. 1986. Statistics in language studies. Cambridge: Cambridge University Press.

Wragg, E.C. 1970. Interaction analysis in the foreign language classroom. Modern Language Journal 54 (2): 116-120.

Xiem, N.V. 1969. The role of explanation in the teaching of the grammar of a foreign language: an experimental study of two techniques. Unpublished thesis, University of California at Los Angeles.