PART I

Evaluation of language education: an overview

Alan Beretta

Introduction

This chapter has two principal aims. The first is to provide a review of previous evaluation studies in foreign language teaching, so that future evaluations may be informed by past experience in the field; the second, to set the evaluation of language education within the broader framework of educational evaluation.

An overview of the history and development of the evaluation of language education since the early 1960s is presented. Although occasional studies were carried out before the 1960s, the past quarter of a century has seen a notable growth in such studies and therefore the period from the 1960s to the present will be the focus of the review. In addition, reference is also made, in the second part of the chapter, to evaluation studies and theory in education more generally. The aim is not to be exhaustive, but to characterise trends and developments in language education evaluation.

To date, very few books have appeared on the evaluation of language teaching programs in general. This compares unfavorably with the general field of educational evaluation, where dozens of titles appear annually in one publishing house alone (Sage). In the language teaching journals, evaluation studies are rarely published which do not focus on the seemingly never-ending 'methods' debate (which had wearied Sweet as long ago as 1899) or on the highly politicised bilingual programs. Yet in the educational and social spheres, specialist evaluation journals proliferate. In TESOL, the major professional organisation for English teachers, there is not even a special interest section on program evaluation. Compare this with the need felt in the American Educational Research Association (AREA) for evaluation to form its own association and hold its own conferences – the American Evaluation Association was formed in 1985 and its major topical interest group is educational evaluation. By comparison, quite clearly, in the field of second language education, there has been little attention given to evaluation.

Throughout the 1970s and into the 1980s, evaluation of language teaching programs has proceeded as if unaware of developments in

educational evaluation. Almost every evaluation study that has been published has used some form of testing and quasi-experimental design regardless, often, of its usefulness in given settings (see Beretta 1986a).

Of course, there have been many evaluations which have been carried out for restricted audiences, and which have therefore not been published. For example, it is known that organisations like the British Council and the British Overseas Development Administration frequently commission applied linguists to undertake evaluations of aid-funded projects. Such evaluations are known to involve the expert visiting the project, often only briefly, and writing a report to the sponsors on their opinions of the value of work to date. However, for-your-eyes-only reports sponsored by government or government institutions are, by their very nature, rarely made available to the language education profession. Thus, it is virtually impossible to know if any of these evaluations are scholarly or disciplined studies. Any evaluation which does not attempt to make public its method of inquiry and its findings can have little to say to an academic audience, so this class of investigation is ignored here.

Some of the evaluations carried out in the language teaching field are published through official channels, and some of these (such as the work carried out at Concordia University by Mackay) bear the hallmarks that a properly trained evaluator would recognise. However, such studies, crucial though they are to the advancement of the field, have not yet been published in the major second language education outlets.

This chapter will briefly trace the short and simple annals of published second language program evaluation and set them within the framework of what has been learnt in the wider field of educational evaluation. The focus will be on the method studies, since the bilingual programs, notably the Canadian Immersion programs, have been documented elsewhere (Genesee 1983, Swain and Lapkin 1982) and until recently were almost invariably investigations solely of product and thus, for the purposes of this book, of limited methodological interest.

25 years of L2 program evaluation

It is not easy to decide where to begin a review of published L2 program evaluation. Certainly, it is barely worth considering much L2 research of any description before 1963, as Agard and Dunkel's (1948) and Carroll's (1963) reviews despondently attest. However, 1963 is a good place to start because it was an auspicious year for educational research in general and for program evaluation in particular. Campbell and Stanley's (1963) monumental treatise on research design appeared that year as did Cronbach's (1963) seminal paper on program evaluation. However,

1963 was also the year that Keating's large-scale evaluation of competing language teaching methods appeared, initiating a disillusionment with evaluation in language education that was quickly compounded by the more famous Scherer and Wertheimer (1964) and Smith (1970) studies, which also compared methods.

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, which found that better results were achieved in classes which did not use a laboratory. Stern (1983:69) recalls that it 'caused a furore'; Freedman (1971:33) dismisses it; Smith warns that 'a careful reading of the study raises serious doubts about the validity of the research' (1970:10). Keating himself makes no claim to the contrary: 'this cannot be considered an experiment in any proper sense' (1963:24). Quite an understatement: there was 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 at all; we know nothing about what happened in control classrooms. Again, Keating is disarmingly 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' (1963:38).

The Scherer and Wertheimer (1964) study, widely known as the Colorado Project, did little to enhance the reputation of evaluation. Scherer and Wertheimer compared audiolingual and cognitive code methods of teaching German and aimed to 'draw some definite scientific conclusions about the relative merits of the two methods' (1964:12). The original plan for a tightly controlled study collapsed and a chapter of accidents ensued: a press leak motivated control students to exercise pressure to join the experimental groups (1964:24); the construction of a new language laboratory was not completed until audiolingual students were to have finished the audio phase of their training; test administration could not be simultaneous for all students because the exam halls were all booked up by other departments. Monitoring of the programs was inadequate and there were insuperable problems with program-fair testing. The results are virtually uninterpretable.

Results failed to show the expected superiority of the audiolingual method in the Pennsylvania Project (Smith 1970). However, these results are, like those of Keating (1963) and Scherer and Wertheimer (1964), extremely difficult to interpret. Distinctions between audiolingual and cognitive code methods were inadequately monitored. There was an attempt at classroom observation, but this did not allow comparison between methods as different schemes were used. Once again, the criterion measures were not program-fair. Once again, a great deal of time and money were spent, expectations were high among the academic

community, and from the very beginning there was never a chance that these expectations would be fulfilled. The Pennsylvania Project signalled the end of the line for large-scale method comparisons, at least in the United States.

It is characteristic of the Pennsylvania Project and other earlier second language evaluation studies that they expected to be able to achieve such tight controls as to be in a position to contribute to theories of language learning. After all, if we knew the best way to promote the learning of languages, many other issues that an evaluation might look at would pale into insignificance. In this sense, it might be considered unfortunate that the Campbell and Stanley paper attracted so much attention from evaluators; its focus on true- and quasi-experimental design, certainly welcome in educational research in general, was less appropriate for evaluation in contexts which typically would be resistant to tight control and theory development. For instance, Scherer and Wertheimer felt able to predict a 'rigidly controlled large-scale scientific experiment which would yield clear-cut data' (1964:12). This was clearly an untenable aim for a study which was to compare audiolingual teaching with cognitive-code, in which the treatments could only be vaguely described and were extremely vaguely monitored, in which neither student nor teacher variables could be controlled, in which all manner of real-world accidents could have occurred (and did), and in which the testing, in my view, at least, could never be program-fair (cf. Beretta 1986b; but also see Bachman 1989 for a more positive view).

By and large, the fall-out from the Pennsylvania Project (Smith 1970) was that those who wished to evaluate methods, concerned at their lack of control in the field, moved to the 'laboratory', investigating aspects of methods over short durations, with at least some variables controlled. For example, Seliger (1975) and Freedman (1976) got around the problem of the teacher variable by replacing teachers with prerecorded lessons, in the process effectively precluding any possibility of external validity. Although the aim was the same – to be able to contribute to theories of language learning – these studies could not have any implications for real world practice as they had been removed from real world classroom settings to more manipulated 'laboratory' inquiries carried out over very short durations (what Eisner 1984:451 has called 'educational commando raids').

To give an idea of the nature of published second language evaluation studies, I have examined thirty-three such studies. The designs of these studies are summarised in Table 1 in terms of duration, number of subjects, whether or not there was some attempt at randomisation, the method being examined, and the strategy adopted (if any) to try to control the teacher variable.

It is evident that the duration of the studies varies considerably, from six lessons or three weeks, to four years. How long an evaluation study should last is obviously an important but difficult issue. Language learning is usually held to be a long-term task, which needs time to be effective. Even if gains are noted in the short term, they may disappear over time. However, the longer the period under study, the greater the contamination from extraneous variables, the greater the risk of drop-out, changes in the project's direction or content, and so on. Yet since evaluation is typically concerned with real-world issues rather than with laboratory effects, studies that show learning or achievement over the long term are going to be more relevant than short-term experiments. It is notable that of the studies examined, only five cover a period of one year or more. The generalisations one can arrive at from an evaluation of a few lessons are obviously very limited.

The studies examined vary greatly with respect to the number of subjects involved, ranging from a total of 21 to a massive 5,000. Most studies involve relatively small numbers of subjects, as can be seen from Table 1. If one takes further account of the design of the studies, however, then the number of subjects in any one group – experimental or control – is even smaller. Wagner and Tilney (1983) and Bushman and Madsen (1976) have only seven subjects per treatment. In cases like these, the small sample sizes call into question the appropriacy of the statistical tests used. It has clearly not been easy for evaluators in language education to gather adequate, representative samples. Nevertheless, the criticism made above of large-scale evaluations like Smith (1970) and Scherer and Wertheimer (1964) stands: the larger the scale of the evaluation, the more things can and do go wrong, and the less the control that can be exercised over events and variables. (This is not to say that all testing comparisons in field settings are wasteful; no methodology can expect to provide the full picture.)

In Table 1, fourth column, 'R' indicates whether any attempt was made to establish the initial equivalence of comparison groups through randomisation or matching procedures. It can be seen that the majority of studies did indeed attempt to match groups, although these attempts did not always result in groups that could be considered equivalent. However, one aim of randomisation is to allow one to generalise results from the sample to the population which, in theory, involves the evaluator sampling from the total population. None of the studies reviewed here would meet this criterion. Indeed, it is unlikely that any real-world evaluation would ever receive sufficient government or financial support to permit such sampling as would allow statistical generalisability. Evaluators have typically had to take whatever subjects they can get. Thus, the pursuit of statistical generalisability in evaluation is unrealistic.

TABLE 1 33 METHOD EVALUATION STUDIES

Study	*Duration*	*No.*	*R/NR*	*Method*	*Control of Teacher Variable*
Keating 1963	1 yr	5,000	NR	AL v CC	None
Scherer and Wertheimer 1964	2 yrs	227	R	AL v CC	Give lesson plans to program teachers
Casey 1968	Ex post facto	50	NR	AL v CC	Questionnaire asking teachers which program they taught
Chastain and Woerdehoff 1968	2 semesters	99	R	AL v CC	None
Smith 1970	4 yrs	1,090	NR	AL v CC	None
Hauptman 1971	3 weeks	69	?	Gr v Sit	None
Mueller 1971	2 semesters	77	NR	AL v CC	None
Levin 1972 (i)	6 lessons	227	R	AL v CC	Recorded lessons
Levin 1972 (ii)	"	104	R	"	Recorded lessons
Levin 1972 (iii)	"	247	R	"	Recorded lessons
Levin 1972 (iv)	"	98	R	"	Recorded lessons
Levin 1972 (v)	"	170	R	"	Recorded lessons
Levin 1972 (vi)	"	57	R	"	Recorded lessons
Levin 1972 (vii)	12 lessons	577	R	"	Recorded lessons
Levin 1972 (viii)	6 lessons	235	R	"	Recorded lessons
Levin 1972 (ix)	"	152	R	"	Recorded lessons
Levin 1972 (x)	10 lessons	125	R	"	Recorded lessons
Asher 1972	32 hrs	37	NR	TPR v Reg	None
Savignon 1972	1 semester	42	NR	Comm v AL	None
Von Elek and Oskarsson 1973	10 lessons	125	R	AL v CC	Recorded lessons
Olsson 1973 (i)	6 lessons	18gps	R	AL v CC	Recorded lessons
Olsson 1973 (ii)	6 lessons	24gps	R	AL v CC	Recorded lessons
Postovsky 1974 (i)	12 weeks	50	R	TPR v Reg	Same teachers for comparison programs
Postovsky 1974 (ii)	12 weeks	48	R	TPR v Reg	Same teachers for comparison programs
Asher *et al.* 1974	1 semester	69	NR	TPR v Reg	None
Gary 1975	22 weeks	50	R	TPR v Reg	Same teacher for comparison programs
Bushman and Madsen 1976	10 hrs	41	NR	Sugg v Reg	Same teachers for comparison programs
Wolfe and Jones 1982	12 weeks	79	R	TPR v Reg	None
Pal 1982	12 lessons	37	R	AL v CC	None
Van Baalen 1983	1 year	80	NR	AL v CC	Questionnaire asking teachers what they taught
Thiele and Scheibner-Herzig 1983	34 lessons	43	NR	TPR v Reg	None
Wagner and Tilney 1983	5 weeks	21	R	Sugg v Reg	Recorded lessons
Beretta and Davies 1985	4 years	341	NR	Comm v Gr	None

Note: AL=Audiolingual, CC=Cognitive Code, TPR=Total Physical Response (and other approaches involving delayed starts in oral production), Sugg=Suggestopedia, Comm=Communicative method, Gr=a grammar-based method, Sit=Situational, and Reg=Regular (unspecified control).

The fifth column of Table 1 shows that all of the studies focused on some method or other: suggestopedia, total physical response, and so on.

Typically, some form of quasi-experimentation was the chosen design, and the aim was to address theoretical problems. The studies were never tightly enough controlled for this task. Indeed they could not be since they took place in real-world classrooms over time. It might reasonably be argued that the designs were quite inappropriate for the questions asked.

The sixth column shows the extent to which the evaluation studies attempted to control the teacher variable (an important element in the standardisation of treatments).

It has been widely recognised that there is a need to control the teacher variable in evaluations that aim to compare programs or to make some theoretical statement. The teacher in Program A might be more highly qualified, more enthusiastic, or may differ in any number of ways from the teacher of Program B. This could offer a rival explanation of results, detracting from the claims of the treatment program. One option is to randomly assign teachers to treatments from a pool large enough to increase confidence that differences are cancelled out. However this has never been achieved even by those who had access to large samples (e.g. Smith 1970, Keating 1963). But even if they had managed it, they could not have controlled for novelty effects: the very newness of a program could produce greater enthusiasm among experimental teachers and students. A second option is to have both programs taught by the same teacher (e.g. Postovsky 1974, Bushman and Madsen 1976), but the teacher may still be more swayed by one than the other with incalculable effects on practice; indeed the Bushman and Madsen paper acknowledges the 'appropriate temperament and excellent teaching skill' of one of the teachers who possessed the right 'philosophical persuasions' for teaching suggestopedic classes (1976:35–7). A third option is to eliminate teachers altogether and replace them with tape-recorded lessons (e.g. Levin 1972). The trouble with this is that it removes the study from the real world and does not permit generalisation beyond the confines of the inquiry. A fourth strategy has been to try to determine what went on in the classroom by asking teachers to fill out questionnaires (e.g Casey 1968, Van Baalen 1983); this strategy is rarely used because of the obvious difficulties associated with self-report.

The notion of standardisation is an important issue in program evaluation. Sometimes there is more variation within programs than between programs, and this has compromised many method comparisons. Attempts to standardise treatments (apart from controlling the teacher variable) are not included in Table 1 because most of the studies make no reference at all to such attempts. There is one exception: the lesson plans used by Scherer and Wertheimer (1964). This can be

considered a partial attempt to standardise the teaching, which along with interviews and class visits may have achieved some regularity. Some researchers mention in passing that they visited some of the classes, and others may have done so without reporting it. But none of the studies attempted to monitor implementation in any principled fashion. Mostly, implementation appears to have been left to chance.

Standardisation of classroom events would be a move in the direction of control, but cannot be accomplished in the field, over time, with real teachers, and all manner of real-world intrusions. If the aim is to claim knowledge about language learning, the impossibility of standardising treatments is one more reason why this aim should be forgotten. However, the need remains to know what happened in the classroom, and so on. The way that programs are implemented is fundamental to evaluation. The most obvious way of gathering this information is through observation, and yet, of the studies in Table 1, only Smith (1970) tried to observe systematically, but he unfortunately used different instruments for the comparison groups, thus precluding comparison.

Table 1 is useful now only for reference and as a historical record. Probably, none of the studies serves as a particularly useful guide to evaluators of language education programs today. First of all, investigations of method have not produced any deeper understanding of the methods involved; second, it has yet to be shown that programs can be subject to tight controls or that comparative testing can be fair. And third, there seems no reason to expect that evaluation can aim primarily to contribute to the advancement of language learning theory. These studies are easy to criticise because we have the benefit of hindsight, but it would be a pity not to learn from them. Anyone who reads them will find that second language program evaluation has only slowly become aware of its existence as a distinct field of inquiry, of the existence of a flourishing evaluation discipline beyond the confines of applied linguistics, and consequently, of any clear sense of its role and direction. It is to evaluation research in the educational and social spheres, which Cronbach *et al.* have called the 'liveliest frontier' of the social sciences (1980:13), that we now turn for sustenance.

What has been learned in educational evaluation?

An explosion of interest in program evaluation occurred in the 1960s. Two reasons are generally offered for this. First, in the wake of the launch of Sputnik in 1957, federal funds in the US were poured into curriculum development in science, mathematics and foreign languages and, eventually, into the evaluation of these programs.

A second reason is that the 'Great Society' reforms of President Johnson in the USA 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' (1977:8).

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*.

Tyler and behavioral objectives

Basically, Tyler's approach, which has since had a tremendous influence on evaluation, involved comparing intended outcomes with actual outcomes. First of all, behavioral 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. Second, the comparison of intended outcomes with actual outcomes does not necessitate the setting up of experimental and control groups. Third, and somewhat problematically, the process of arriving at behavioral 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 *topics* 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 *et al.* 1980: 173–4)

The problem with confining evaluation to behavioral 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.

To take another example: the Bangalore Project (Prabhu 1987) rested upon an incubation hypothesis; 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 two 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: 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: 103)