PROCESS AND PRODUCT IN ESL PROGRAM EVALUATION

Michael H. Long

1. Introduction

This paper seeks to make a case for 'process' evaluation of ESL programs. It does not advocate process evaluation alone, however, but rather, as an essential supplement to the usual product evaluation of those programs' most important outcome, ESL development. The process/product distinction is compared with that between formative and summative evaluation (Scriven, 1967), but is not intended to replace it. The two reflect different, not competing, perspectives. The final section outlines the role in process evaluation of classroom-centered research.

2. Product evaluation

Most program evaluations are product-oriented. That is, they focus (quite reasonably) on what a program produces, chiefly in terms of student learning, but sometimes also in terms of changes it brings about in teachers' and students' attitudes, students' self-concept, related intellectual skills, and the like. Thus, most product evaluations set out to answer one or both of the following questions:

(1) Does program X work?

(2) Does program X work better than program Y?

Question (1) is concerned with a program's absolute effectiveness, question (2) with the relative utility of one program compared with another. (See Long (1983) for a recent review of evaluations of both types.) Question (1) is often asked when a new program has been established. For instance, can graduates of a new EAP program follow lectures in English and extract information from English textbooks efficiently enough for them to register in university credit courses in their specializations? Question (2) is more typical when an existing program is undergoing some curricular changes. An example would be the introduction of a notional-functional track in a program with a hitherto structurally based curriculum.

In order to answer question (1), as is well known by now, it is not enough simply to pretest entering students, put them through the program, and then test them again to see if they have reached criterion level. Even if all students score 100% on the post-test, one cannot conclude that program X works. One wants to know, after all, whether the improvement was achieved as a result of program X, as distinct from while enrolled in program X or, worse, despite program X. In other words, answering question (1) means establishing a causal relationship between program X and ESL development. This, in turn, means employing a true experimental design in the evaluation: minimally, one group of students doing program X, another group of students, equivalent in all respects to the first, acting as controls, with both groups having been formed by random assignment from an initial pool (see Figure 1).

Figure 1 about here

Even this, of course, is insufficient basis for a claim that program X works, as should soon become clear.

In order to answer question (2), a modification of this design can be used, wherein the rival curriculum, program Y, substitutes for the control group's filler activity (see Figure 2).

Figure 2 about here

Use of three groups, one doing program X, one doing program Y, and a control group, each group again formed by random assignment, allows the evaluation to answer questions (1) and (2), and so is more cost-effective (see Figure 3). One would like to think, of course, that question (1) will already have been answered in the affirmative before the curricular innovation leading to question (2) makes a second evaluation necessary. In practice, however, this is seldom the case. Rival curricula are often in competition before the claims of either to do a job have been verified. Witness the waves of language teaching methods and approaches to syllabus design presently buffeting the good ship TESOL.

Figure 3 about here

There are several well known threats to the internal validity of studies utilizing the classical experimental approach to program evaluation outlined above. Briefly, these include:

1. History. Something happens during the course of a study, the effects of which are not controlled for and which could constitute an explanation for the results obtained, either alone or in combination with the effects of the program. For example, students in the treatment group (program X) might make friends with English speakers, and improve their English simply by conversing with them outside the classroom, or in this way and through program X.

2. Maturation. In programs lasting several weeks or months, changes may occur in the students which improve their post-test scores, yet have nothing to do with program X. For example, through residence in an English-speaking environment, students may develop more positive attitudes to the target language and its speakers. This may translate into higher motivation to learn, and this into higher achievement, independent of the (supposed) benefits of instruction.

3. Testing. The use of a pre-test can have two undesirable side-effects. First, the pre-test can sensitize students to the subject-matter being tested, and alert them to this when doing the post-test. Second, doing the pre-test can help students learn the material being tested, and so lead to improved performance independent of the effects of instruction. For example, doing the items in a discrete point grammar test is not unlike doing grammar exercises on the points tested. Doing the pre-test is an additional practice opportunity which may help students improve their scores when tested again.

4. Instrumentation. Flaws in the testing instruments themselves can determine the outcome of a study, as can inconsistencies in their administration. Thus, if the tests are unreliable and/or invalid, it will be impossible to interpret students' test scores at all. If the tests are valid, but administered under different conditions, the same may be true. Suppose, for example, that more time is given for the post-test. Student performance may be expected to improve. Lastly, two tests may be valid, and administered under identical conditions, but the post-test turn out not to be an equivalent form of the pre-test, e.g., through being easier.

5. Selection. Students may be selected for one group in a study who differ in some important way from students in the comparison group(s). For example, unknown to the evaluator, students in one of the groups may be more motivated or more intelligent than students in the other group(s).

6. Mortality. Students may drop out of a program during the course of a study, disappearing from one or all the groups involved. This vitiates the findings in cases where the dropouts are not systematically accounted for in the analyses. Suppose, for example, that program X is not better than program Y, but is intellectually more challenging. Several of the weaker students drop out of program X, with the result that the average ability/ proficiency level of the students remaining in the program is higher than that of students in program Y. The average post-test scores for the élite group of survivors in program X should be higher than those of the more heterogeneous group of program Y students, independent of the effects of the two programs.

If any one of these six threats to internal validity becomes a reality, an evaluation is in potentially serious trouble. Most can easily be avoided, however, and some can be rectified during the analysis stages of a study. (For further discussion and details, see, e.g., Genesee, 1983; Hatch and Farhady, 1982; Shavelson, 1981; Swain, 1978; Tucker and Cziko, 1978.)

3. Limitations of product evaluation

Suppose one is sure that none of the above six factors has invalidated an evaluation. Can question (1) or (2) be answered with confidence through a well executed product evaluation?

Unfortunately not. Some hypothetical examples may clarify why this is so. The following are just some of the possible outcomes of product evaluations designed to answer questions (1) and (2), together with a few of the many possible (hidden) explanations for those outcomes. Many such explanations have actually been uncovered in real evaluation studies, though not necessarily of ESL programs. (See, e.g., Swaffer, Arens and Morgan, 1982; Tyler, 1975.) All of them have probably occurred at some time, but gone undetected. The examples are presented in tabular form (see Figure 4) in an attempt to promote readability.

Figure 4 about here

The items in Figure 4 are just a few of the many possible outcomes, and an even smaller selection of their possible (hidden) explanations. They should make it clear that an exclusively product-oriented evaluation--even one uncontaminated by any of the threats to internal validity--is inadequate. By focusing on the product of a program, while ignoring the process by which that product came about, it is in serious danger of providing false information. Product evaluations cannot distinguish among the many possible explanations for the results they obtain.

Does this mean that product evaluation should be abandoned in favor of process evaluation? Absolutely not. Product evaluation is essential. A process evaluation which ensured that a certain set of desired classroom processes did in fact obtain in the classrooms under study would still need to determine that these processes actually produced the anticipated results. What is needed is both process and product evaluation. The point is, however, that, while necessary, product evaluation alone is not sufficient.

4. Process evaluation

Thus far, the reader has been left to arrive at an intuitive understanding of what is meant by 'process' evaluation. It is time to be a little more explicit. There are, after all, a host of verbal and non-verbal behaviors by teachers and students which contribute to ESL classroom processes. Further, these are multifaceted, being analyzable in pedagogic, linguistic, psycholinguistic and sociolinguistic terms, among others. And then there are language-learning materials, which can also be analyzed in a variety of ways. Clearly, to choose rationally among all these possibilities, one needs a theory. Since ESL development is what is at issue, this will obviously mean a theory of (second) language acquisition. Thus, by 'process evaluation' is meant the systematic observation of classroom behavior with reference to the theory of (second) language development which underlies the program being evaluated. An example is in order.

An established English Language Institute in the USA has for some years used structural-situational language teaching materials taught via a modified audio-lingual method. Recently "converted" to a radically different set of beliefs about how adults learn a second language, the director and a group of her teachers decide to try out the Natural Approach in two of their intermediate classes, and to compare the results with those obtained in two of their regular audio-lingual classes at the same level. They set up their product evaluation as outlined in Figure 2. Aware, however, that ESL teachers rarely stick to a single "method" over time, and aware, too, that even apparently very different "methods" often overlap at the classroom level in terms of some of the activities students engage in, they decide to take two precautions. First, they hold a pre-session workshop in which the teachers involved are thoroughly familiarized with the classroom procedures they are supposed to follow in each program. Each group of teachers agrees to stick to these for the duration of the study (one semester). Second, to ensure that this is in fact done, and to make the product evaluation findings interpretable, they decide to do a process evaluation.

Unfortunately, the ELI concerned has just suffered its third budget cut in as many years in order to help the university expand its business administration program, (40% of whose students are non-native speakers, incidentally), and so is not blessed with such luxuries (for an ELI) as VTR equipment. It is therefore decided to collect process data by simple audio-taping. Every two weeks, one lesson in each class will be recorded and transcribed. Given a 16-week semester, the data base for the process evaluation will comprise transcripts of eight 50-minute lessons per class, for four classes--a total of 32 transcripts. After the transcripts have been coded for certain features, and inter-rater reliability checks conducted (see, e.g., Frick and Semmel, 1978, for details), the relative frequencies of these phenomena in the two kinds of classes will be compared. This will enable the evaluators to ensure (1) that the two programs were observably different from each other, not just on paper, but in the classroom, and (2) that the observed behaviors in particular classes corresponded to those required by the program each class was assigned to!

At approximately five minutes per minute of tape, transcription for this study will take upwards of 32 (lessons) by 50 minutes by five minutes--a total of about 116 hours, not counting verification of transcripts. The coding and quantitative analyses will take roughly another 40 hours, making a grand total of about 150 hours. The ELI director applies to her university for a small R & D grant ($900) with which to pay for some graduate student assistance with the transcription and coding (at $6 per hour), plus the cassette tapes. The request is turned down, although the refusal letter encourages her to pursue what is "clearly a most commendable project". (Unfortunately, it was just beaten out by a bid from computer science for additional cleaning staff for one of their new computers.) The ELI has yet to have one of its grant applications funded, and so the evaluation team is not deterred. The members decide to reduce the transcription time by half through sampling from the tapes, and to do the work themselves on a voluntary basis. (They are all non-unionized, part-time university employees, after all, and so have plenty of spare time in the afternoons.)

What the evaluators look for in the transcripts is determined by the nature of the two programs in question. The team draws up a list of the main features of Audio-lingualism and the Natural Approach, and compare the two. There are obvious differences. They note, among other things, that the former advocates (1) structural grading, (2) immediate, forced oral production by students, (3) avoidance and correction of errors, i.e., a focus on form, (4) both mechanical and meaningful language practice, chiefly through the memorization of short dialogs built around basic sentence patterns, and (5) large doses of drillwork. The Natural Approach, on the other hand, rejects all five (see, e.g., Krashen, 1982).

The next step is to choose categories of classroom behavior (preferably frequent, low inference categories) which will distinguish the two programs at the classroom level. The team opts for just two of the five: (2) error correction and (4) level of language use. While all five features could probably be operationalized, some might be problematic. For example, while the Natural Approach rejects structural grading, there is a certain amount of "natural" structural grading in teacher speech, which is (roughly) tuned to students' second language proficiency by the effort to communicate (Gaies, 1977).

Transcripts will be coded in two ways. First, morphological and syntactic errors will be identified, and the teacher's speaking turn following each error will be coded for the absence/presence of a "correcting" move of some kind. Second, as a simple index of the non/communicative nature of classroom language use, all utterances in the teachers' speech which function as questions will be identified, and then classified into one of two categories: display questions ('Are you a student?') and referential questions ('Has anyone seen Maria's bag?').

In the manner described, a simple process evaluation has been designed. While admittedly crude, it will probably suffice for the purposes of this evaluation. The reader familiar with the TESOL literature of the past few years will have noticed that the design was influenced by a few of the findings of classroom-centered research during that period. Before proceeding to outline the role this work can play more fully, some differences should be noted between what is here being called 'process' and 'product' evaluation, and Scriven's terms, '_formative_' and '_summative_' evaluation (Scriven, 1967).

5. A comparison of process/product and formative/summative evaluation

5.1 _Formative and summative evaluation_. As is by now well known, formative and summative evaluation differ in at least three ways: in focus, timing and purpose.

5.1.1 _Focus_. Formative evaluations typically look at such factors as teachers' and students' attitudes to a curricular innovation, or at the usability of new instructional materials as they are tried out in the classroom for the first time. Summative evaluations, on the other hand, generally measure student achievement in the ways described under product evaluation, and also such matters as cost-effectiveness.

5.1.2 _Timing_. Formative and summative evaluations differ in the importance attached to their timing (Levy, 1977, p. 12). Formative evaluations assess the strengths and limitations of a new program as it is developed and implemented. Summative evaluations are carried out after the development and implementation process is complete.

5.1.3 _Purpose_. The purpose of the two types of evaluation differ. Information obtained from formative evaluations about such matters as the transparency/opacity of new instructional materials to teachers and students, or about unforeseen cultural problems the materials give rise to, is sought by program developers with a view to modifying a program as it is being implemented, or formed (hence, 'formative'). Summative evaluations attempt to summarize (hence, 'summative') the results of a program, once implemented. Now that teachers have taught the new program, and students have passed through it, student learning, teacher and student attitudes and cost-effectiveness can be judged. The purpose of such an assessment is usually to determine whether or not the program should be continued.

5.2 Summative and product evaluations. It can now be seen that summative and product evaluations will sometimes, but now always, be the same. They tend to differ most frequently in two areas: scope and content.

5.2.1 Scope. Summative evaluations are typically broader in scope, often assessing attitudinal or cost issues, for example, as well as student achievement. Product evaluations, on the other hand, tend to be more restricted, focusing on student achievement as the most important outcome issue.

5.2.2 Content. Note, however, that by no means all summative evaluations address second language development issues at all. Many bilingual education evaluations, for example, have concentrated on such issues as students' self-concept and students' attitudes to the native and second language and/or culture. In such cases, there is a qualitative, not just a quantitative difference involved.

5.3 Formative and process evaluations. The difference between formative and process evaluations is more obvious and, it might be claimed, more important. The difference between the two may include focus, theoretical motivation, timing and purpose.

5.3.1 Focus. While some formative evaluations in the past have utilized classroom observational techniques, among other data-gathering devices (see, e.g., Yolonde, 1977), this has been the exception, not the rule. Gathering classroom process data is the essence of process evaluation, on the other hand.

5.3.2 Theoretical motivation. The kinds of classroom processes examined in those formative evaluations that have considered them at all have almost always (always?) been of pedagogical, not psycholinguistic interest. Thus, formative evaluations have collected data on such phenomena as classroom organization (lecture mode, group or individual activity, whole-class discussion, etc.), on the pedagogic function of utterances (instruction, suggestion, lecture, praise, express opinion, etc.), or on the amount of time spent on different content areas via different modalities. While no doubt relevant to curricular innovations in content areas, few of these are analyses which could readily be motivated by any current theory of second language development. One may say, therefore, that formative and process evaluations differ in theoretical motivation.

5.3.3 Timing. The timing of formative evaluation has already been identified as during the development and implementation phase of a new program. Process evaluation, by way of contrast, will be carried out on established (fully developed and implemented) programs.

5.3.4 Purpose. Whereas the purpose of formative evaluation is just that, 'formative', process evaluation, as described earlier, seeks to provide explanations for the findings of product evaluations.

5.4 Complementary roles of the four types of evaluation.

The comparison of process/product and formative/summative evaluations shows them to differ in a variety of ways. There is no suggestion, however, that one should replace the other. Rather, they reflect different perspectives, different goals that evaluations may have. Some of the differences are made explicit by the terms themselves. Others are left implicit. Thus, implicit in the process/product distinction is a sense that language learning classrooms differ in some fundamental ways from content classrooms --do differ, not necessarily should differ--and that these differences need to be reflected in the ways they are evaluated.

The root cause of the language learning/content classroom differences, of course, is the fact that in most second language lessons, language is both the vehicle and object of instruction. (Hence, the great interest to TESOL of immersion education and the current research on "sheltered content classes". See, e.g., Wesche, 1982.) Such linguistic and psycholinguistic phenomena as modeling, error, correction, input, conversation, simplicity, saliency and frequency, for example, have relevance in some areas of content curricula. They have special significance, and often special connotations, in second language classrooms, however. They are just some of the constructs and concepts which figure in modern theories of (second) language acquisition. They can easily find a place in process evaluations of the kind outlined in this paper, but are unlikely to be addressed in product, formative or summative studies.

6. Process evaluation and classroom-centered research

Classroom-centered research (CCR) was noted earlier as a useful source of ideas in the design of process evaluations. This should not be surprising, for while most CCR to date has been descriptive, not evaluative, the object of study has been exactly what is being proposed as suitable for process studies in an evaluation context.

Much CCR of the last decade grew out of disillusionment with large-scale, global "comparative method" studies of the 1960s. Studies such as the Pennsylvania project (Smith, 1970) and Colorado project (Scherer and Wertheimer, 1964) attempted to compare grammar translation and audiolingualism, or audiolingualism and "cognitive" methods of instruction, in much the same way as the product evaluations of individual programs described in this paper. The comparative method studies differed in size and duration, however. They lasted up to three years, and attempted to follow large numbers of intact classes (and their teachers) assigned to one or other of the "methods", employing only a limited number of rather superficial class observations or none at all. The results were generally inconclusive, and were anyway difficult to interpret for precisely the same reasons that product evaluations (alone) have been criticized here. Reviewers at the time complained of the lack of verification that the methods were adhered to, suspecting a large amount of overlap at the classroom level. (See, e.g., Freedman, 1975; Levin, 1972.) They further doubted whether the various "methods" were clearly enough defined in the teachers' minds.

Findings of a recent study (Swaffer, Arens and Morgan, 1982) show how justified such fears may have been. Before participating in a smaller study of two approaches to the teaching of German, teachers received careful training in the procedures each group was to follow. After the study was completed, Swaffer et al. debriefed the teachers, in part seeking to determine the degree to which the two approaches were now clear in the subjects' minds. The confusion they uncovered led the investigators to the following conclusion:

> . . . defining methodologies in terms of the characteristic activities has led to distinctions which are only ostensible, not real, i.e., not confirmable in classroom practice.
> (Swaffer _et_ _al_., _op_._cit_., p. 32)

Similarly depressing results have been obtained in two classroom-centered studies of the effects on classroom language use of the introduction of (supposedly) different types of teaching materials. Having written some new notional-functional ESP materials for a university in Iran, Phillips and Shettlesworth (1975) decided to compare the discourse engendered by their materials and that in lessons using the structural-situational materials the new ones were intended to replace. After studying transcripts of lessons in the two types of classrooms, the researchers concluded:

> (O)ur analysis of the samples of discourse engendered by these courses leads us to the conclusion that they all tend to structure the lesson in a similar manner; this suggests, therefore, that the ESP courses at least are failing in their intent.
> (_opus_ _cit_, p. 7)

Another study of this sort, this time conducted in Mexico, found that some newly produced "communicative" language teaching materials affected classroom discourse only when the materials were utilized in conjunction with small group work (Long, Adams, McLean and Castaños, 1976). The materials alone had a negligible impact on the kinds of speaking opportunities students received, whether these were analyzed in pedagogical, functional or social-interactional terms.

Most recently, two additional studies have looked at conversational patterns in ESL classrooms when two other variables are manipulated. Long and Sato (1983) compared language use in lessons taught by teachers recently trained in "communicative" approaches in three major MA in TESL programs in the USA with that of native speakers conversing with non-natives of the same ESL proficiency outside classrooms. Striking differences were found in the quality of language use in the two settings, the ESL lessons consisting predominantly of the same mechanical and meaningful (not communicative) language use (chiefly question-and-answer drills) documented in pre-"communicative" era trainees. (The informal native/non-native conversations, on the other hand, consisted entirely of genuine communication.) Subsequently, Pica and Long (1982) confirmed these findings in a comparison of experienced and inexperienced ESL teachers teaching the same classes of students.

Results such as these confirm the importance of looking at the <u>process</u> of second language learning in classrooms before making any assumptions about the independence of two programs in an evaluation study. They are the kind of findings that have given additional impetus to a growing number of researchers in their resolve to give due weight to the language learning process in their work, rather than to focus exclusively on the product of acquisition. This is true not only in classroom studies, but of research on naturalistic second language acquisition (see, e.g., Pica, 1982) and of psycholinguistically motivated approaches to syllabus design (see, e.g., Pienemann, 1983).

While still in its youth, if not infancy, CCR has already accumulated a substantial body of knowledge about what actually goes on in ESL classrooms, as opposed to what is believed to go on, and as distinct from what writers on TESL methods tell us ought to go on. Topics investigated include teacher feedback on learner error, teacher questions, turn-taking systems, language use in lockstep and small group work, simplification in teacher speech, vocabulary explanation, interlanguage talk and ethnic styles in classroom discourse. Several reviews of findings are now available (see, e.g., Allwright, 1983; Bailey, in press; Gaies, 1983). A lot has also been learned about methodological issues in conducting such research (see, e.g., Chaudron, 1983; Long, 1980), and operationalized definitions of many relevant process variables can be found in the original research reports, several of which could provide almost ready-made check-lists for evaluation studies. Their utility in such studies will only be appreciated, however, when ESL program evaluators broaden their focus to include process, not just product evaluation.

Figure 1: Simplest (true experimental) design for a product evaluation of the absolute effectiveness of a program[1]

$$\begin{array}{c} R\ (O_1)\ X\ O_2 \\ \hline R\ (O_1)\ \emptyset\ O_2 \end{array}$$

Figure 2: Simplest (true experimental) design for a product evaluation of the relative utility of two programs

$$\begin{array}{c} R\ (O_1)\ X\ O_2 \\ \hline R\ (O_1)\ Y\ O_2 \end{array}$$

Figure 3: Simplest (true experimental) design for a product evaluation of the absolute effectiveness and relative utility of two programs

$$\begin{array}{c} R\ (O_1)\ X\ O_2 \\ \hline R\ (O_1)\ Y\ O_2 \\ \hline R\ (O_1)\ \emptyset\ O_2 \end{array}$$

[1] In figures 1-3, R = group formed by random assignment; O_1 = pre-test (first observation); X = program X (treatment); Y = program Y (treatment); Ø = filler activity for control group; O_2 = post-test (second observation).

Figure 4: Some of the possible outcomes of product evaluations, and some of the possible explanations for those outcomes (Or: A rationale for process evaluations)

A. Question (1): Does program X work?

Outcome 1: Program X students (Ss) pass O_2; control Ss do not.

Explanation 1a: Program X works, and teachers (Ts) and Ss did X.

Explanation 1b: Program X does not work, but Ts and/or Ss in X did A, not X, and A works.

Explanation 1c: Any kind of program would work. Ts and/or Ss in X did A, not X, and A works.

Outcome 2: Program X Ss and control Ss do equally well/badly (no difference between groups)

Explanation 2a: Program X works, but Ts and/or Ss in X did B, not X, and B does not work.

Explanation 2b: Program X does not work, and Ts and Ss did X.

Explanation 2c: Program X does not work, but Ts and/or Ss in X did B, not X, and B does not work, either.

B. Question (2): Does program X work better than program Y?

Outcome 1: Program X Ss score higher (or improve more) than program Y Ss.

Explanation 1a: X works better than Y, and program X and Y Ts and Ss did X and Y, respectively.

Explanation 1b: X works better than A, and program Y Ts and/or Ss did A, not Y.

Explanation 1c: There is no difference between X and Y, but program Y Ts and/or Ss did A, not Y, and X works better than A.

Explanation 1d: Y is actually better than X. But program X Ts and/or Ss did B, not X. Program Y Ts and/or Ss did C, not Y. B is better than C.

Outcome 2: Program X Ss and program Y Ss do equally well/badly (no difference between groups)

Explanation 2a: X works better than Y, but program X Ts and/or Ss did A, not X, and there is no difference between A and Y.

Explanation 2b: Y works better than X, but program Y Ts and/or Ss did B, not Y, and there is no difference between X and B.

Explanation 2c: X works better than Y, but Ts and/or Ss in both programs mixed X and Y in their classes.

REFERENCES

Allwright, R. L. 1983. Classroom-centered research on language teaching and learning: a brief historical overview. TESOL Quarterly 17, 2, June, 191-204.

Bailey, K. M. In press. Classroom-centered research on language teaching and learning. In M. Celce-Murcia (ed.), Essays for Language Teachers, Rowley, Mass.: Newbury House.

Chaudron, C. 1983. The use of metalinguistic judgments in classroom research. Paper presented at the 17th annual TESOL Conference, Toronto, Canada, March.

Freedman, E. S. 1975. Experimentation into foreign language teaching methodology. Paper presented at a meeting of the British Association for Applied Linguistics, York, September.

Frick, T. and M. I. Semmel. 1978. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research 48, 1, 157-184.

Gaies, S. J. 1977. The nature of linguistic input in formal second language learning: linguistic and communicative strategies in ESL teachers' classroom language. In H. D. Brown, C. A. Yorio and R. H. Crymes (eds.), On TESOL '77, 204-212. Washington, D.C.: TESOL.

Gaies, S. J. 1983. The investigation of language classroom processes. TESOL Quarterly 17, 2, June, 205-217.

Genesee, F. 1983. Bilingual education of majority-language children: the immersion experiments in review. Applied Psycholinguistics 4, 1, March, 1-46.

Hatch, E. M. and H. Farhady. 1982. Research Design and Statistics for applied linguists. Rowley, Mass.: Newbury House.

Krashen, S. D. 1982. Principles and Practice in Second Language Acquisition. New York: Pergamon.

Lewy, A. 1977. The nature of curriculum evaluation. In A. Lewy (ed.), 3-33.

Lewy, A. (ed.) 1977. Handbook of Curriculum Evaluation. New York: Longman.

Long, M. H. 1980. Inside the "black box": methodological issues in classroom research on language teaching. Language Learning 30, 1, 1-42.

Long, M. H. 1983. Does second language instruction make a difference? A review of research. TESOL Quarterly 17, 3, September, 359-382.

Long, M. H., L. Adams, M. McLean, and F. Castaños. 1976. Doing things with words: verbal interaction in lockstep and small group classroom situations. In J. Fanselow and R. Crymes (eds.), On TESOL '76, 137-153. Washington, D.C.: TESOL.

Long, M. H. and C. J. Sato. 1983. Classroom foreigner talk discourse: forms and functions of teachers' questions. In H. W. Seliger and M. H. Long (eds.), Classroom-oriented research on second language acquisition. Rowley, Mass.: Newbury House, 268-286.

Phillips, M. and C. Shettlesworth. 1975. Questions in the design and implementation of courses in English for specialized purposes. Paper presented at the 4th AILA Conference, Stuttgart.

Pica, T. 1982. Second language acquisition in different language contexts. Unpublished Ph.D. dissertation. Philadelphia: University of Pennsylvania.

Pica, T. and M. H. Long. 1982. The linguistic and conversational performance of experienced and inexperienced ESL teachers. Paper presented at the 16th annual TESOL Conference, Honolulu, HI. May.

Pienemann, M. In press. Learnability and syllabus construction. To appear in K. Hyltenstam and M. Pienemann (eds.), Instructional and Social Implications of L2 Acquisition Research. London: Multilingual Matters, 1984.

Scherer, A. C. and M. Wertheimer. 1964. A Psycholinguistic Experiment in Foreign Language Teaching. New York: McGraw-Hill.

Scriven, M. 1967. The methodology of evaluation. In R. W. Tyler (ed.), Perspectives on Curriculum Evaluation. Chicago: Rand, McNally, 39-83.

Shavelson, R. J. 1981. Statistical Reasoning for Behavioral Sciences. Boston: Allyn and Bacon.

Smith, P. D. Jr. 1970. A Comparison of the Cognitive and Audiolingual Approaches to Foreign Language Instruction. The Pennsylvania Foreign Language Project. Philadelphia, Pa.: The Center for Curriculum Development.

Swaffer, J. K., K. Arens and M. Morgan. 1982. Teacher classroom practices: redefining method as task hierarchy. Modern Language Journal 66, 1, 24-33.

Swain, M. 1978. School reform through bilingual education: problems and some solutions in evaluating programs. Comparative Education Review, October, 420-433.

Tucker, G. R. and G. A. Cziko. 1978. The role of evaluation in bilingual education. In J. E. Alatis (ed.), International Dimensions of Bilingual Education. 423-446. Washington, D.C.: Georgetown University Press.

Tyler, R. W. 1975. The Activity School Project. Viewpoint 51, 12-31.

Wesche, M. 1983. Report of research on "sheltered subject-matter classes." Paper presented at the 17th annual TESOL Conference, Toronto, Canada, March.

Yolonde, E. A. 1977. Observational techniques. In A Lewy (ed.), 189-209.