VALIDATION IN SECOND LANGUAGE CLASSROOM RESEARCH
THE ROLE OF OBSERVATION[1]

CRAIG CHAUDRON
University of Hawai'i at Manoa

Til Claus

Sommernatslyset svindende i bålets
gnister drog os sammen,
drøftede forskning
og baggrund;
fra modsatte retninger,
begge ved sit vejkryds.
Gid du kunne fortsætte med os.

One of the impressive features of Claus Færch's work was his exemplification of not speculating on principles of language learning and teaching without having closely studied their substantive processes and products. His paper on "Rules of Thumb..." (1986) is a clear case of derivation through classroom observation of a general pattern for teacher-formulated rule use in classrooms, which then opened up a large area for further research into the "real" goings on of grammar teaching. Claus was fully aware, however, that his induction of steps in rule presentation might not, in the end, prove to be the critical features of teacher and student interaction in grammar learning. Yet they provided him with a framework within which he could place his further observations. The process of research inherently involves a dynamic validation, application, and reevaluation of such constructs through empirical testing and theoretical restructuring (cf. Chaudron 1986a).

The purpose of this paper is to explore more precisely how

[1] Paper to appear in *Foreign Language Pedagogy Research: A Commemorative Volume for Claus Færch*. Edited by R. Phillipson, et al. Multilingual Matters.

I am indebted to Graham Crookes for valuable assistance in obtaining sources for this paper, and for his helpful comments on a first draft of the paper. Thanks also to Merrill Swain for suggestions.

University of Hawai'i Working Papers in ESL, Vol. 7, No. 2, December 1988, pp. 1–15.

observational analyses of classroom interaction can be validated, and further, how the validation of claims about instructional variables (such as the effectiveness of programs, teaching methods, syllabus changes, materials, rule presentations, and so on) depends on the application of valid observational analyses. Validity, which has many aspects but refers in essence to the determination of the "truth" of an analysis or theory, is a fundamental goal in researchers' efforts to understand and predict language learning and teaching outcomes. The paper will first briefly describe the place of observation in research validation, then show the applicability of validation in second language classroom research with respect to different methodological orientations, then illustrate three different approaches to validation of instructional research by means of observation.

Campbell & Stanley (1972) proposed the general research considerations of "internal" and "external" validity. These refer respectively to the truth of observations within a study, and to the generalizability of the observations or findings across studies. Measurement systems (such as tests) and observational procedures, which are subject to the form of internal validity known as "instrument validity," play a vital role in ensuring overall internal validity, namely by documenting that relevant treatments and processes in fact occur. Observations of a language classroom, whether by means of planned schemas or post hoc characterizations and discourse analysis, must undergo an evaluation of their reliability as descriptions (by means of intra- and interobserver consistency checks). This reliability assessment is the initial step in instrument validation. However, the validity of such observational descriptions as constructs relevant to the research questions can only be fully attained if the observations and summary findings of the study are shown to hold in more general ways (external validity). Such validation is accomplished through rigorous application of sampling procedures and design principles. In general, the same or similar observational analyses must be applied to new contexts and populations. (See current treatments of many of these issues in Frick & Semmel 1978, Brinberg & Kidder 1982, Brinberg & McGrath 1982, 1985, Cone 1982, LeCompte & Goetz 1982, Folger, Hewes & Poole 1984, Hoge 1985, and Poole & McPhee 1985.)

Methodological Orientations

Researchers can describe classroom events according to various theoretical perspectives, which lead to different methodological orientations to observational analysis. These orientations might be broadly characterized as, for example, "classification" or "process" (van Lier 1984), "systematic" or "interpretive" (Edwards & Westgate 1987), "interaction analysis," "discourse analysis," or "ethnographic" (Chaudron 1988; see earlier discussion in Long 1980). The principal distinction among these rests on the degree to which an exhaustive and structured set of categories of behavior are used to describe the interaction.

Classification, systematic, and interaction/discourse analysis perspectives use precisely defined observational categories organized in structured systems (as in Moskowitz' FLint system 1976, Sinclair & Coulthard 1975, Fanselow's FOCUS 1977, Allen, Fröhlich & Spada 1984, and others related to these). Process-oriented, interpretive, and ethnographic perspectives adopt context-dependent, location-specific descriptions, often only after observation has begun (as in various applications in Allwright 1975, Trueba, Guthrie & Au 1981, van Lier 1982, and Bailey 1983). As is clear from the general educational and psychological literature on validity cited above, and as I have discussed on other occasions (1986a, b, 1988:Chapter 2), such descriptions, like any measurement instruments, must be evaluated for their reliability first, and then for other forms of validity.[2] This is so regardless of the theoretical perspective taken, as LeCompte & Goetz (1982) demonstrate quite clearly in discussing reliability and validity in ethnographic research.

As an illustration of this rationale for validation, McCutcheon's (1981) exposition of validity in educational research is worth reiterating and applying to a recent ethnographic study (Enright 1984). While McCutcheon was especially concerned with qualitative, process-oriented research, (as in ethnography) her argument applies equally well to the classification-minded, quantitative researcher who might, for example, adopt the set of descriptors in Sinclair and Coulthard's (1975) type of hierarchical discourse analytical system

[2] It is rare for L2 classroom researchers to assess the reliability of their instruments, much less other forms of validity. For more details on reliability assessment in observation systems, see Frick & Semmel (1978), LeCompte & Goetz (1982), Page & Iwata (1986), and Chaudron (1988). For some exceptions to the rule of failure to determine reliability in L2 research see especially Mitchell, Parkinson, & Johnstone (1981), and Chaudron (1988:24n.).

(as recently done in Ramirez 1988). While avoiding use of the term "validation," McCutcheon is clearly proposing a basis for validating such research. She claims that interpretation of observations comprises three types: 1) "the forming or construction of patterns," 2) the discernment of "the social meaning of events," and 3) "the relating of particulars of the setting to external considerations," such as theories or other events. In order to judge any of these three types of interpretation, however, McCutcheon further points out that either the interpreter or others must evaluate them on the basis of a) the LOGIC or argumentation, b) the sufficiency of EVIDENCE, c) the agreement or CONSISTENCY with other evidence, and d) the SIGNIFICANCE of the analysis (in terms of theoretical additions or predictive value, etc.). Furthermore, in assuming an audience for their interpretations, researchers also necessarily expect *inter*subjective (i.e., "objective," in the sense of "public") understanding of their descriptions, and some degree of generalizability of the descriptions to other situations. These various methods of evaluating interpretations are commonsense expressions of the distinct ways in which research observations are validated.

For example, Enright's (1984) predominantly ethnographic study uses the concept "participant structures," a relatively low-inference unit of observation applying to the changes in "configurations of concerted action," as a basis for his analysis of the differential choices available to teachers. By varying aspects of participant structure, teachers can engender more or less student turn-taking. Regardless of Enright's research question, his observations and categorizations are still subject to the constraints McCutcheon referred to if we are to regard them as valid.

I will refer to McCutcheon's points a) through d) to illustrate how Enright's analysis needs validation. First, there must be an internal LOGIC (point a)) that systematically interrelates the different participant structures (a type of "construct" validity—to be elaborated on below). In order to ensure that these constructs are viable, there is a need for the prior evaluation of instrument validity in the form of interobserver reliability: the changes in dimensions of actors, topics, etc. that together constitute different participant structures must be identifiable and recognized by independent observers (intersubjectively). These steps are essentially equivalent to determining the "content" validity of the observations, as one would do in verifying that items on a test represent the skills or knowledge individuals are being tested on. If a

new observer is actually presented with the data on which the researcher's claims are based, interoberserver reliability is tested. If the observer only evaluates the reasonableness of the interpretations and generalized observations, a limited sort of content validation is conducted.

Skipping over point b) for the moment and considering point c), observations of participant structures must also be CONSISTENT with other behavior associated with them ("concurrent" validity), which Enright attempts to demonstrate (see next paragraph). Finally, with regard to point d), the analysis must prove to have some bearing on further findings with these or similar teachers and contexts (generalizability), and should broadly have consequences for better understanding or control over teaching and learning in such contexts.

Enright's primary concern is in fact to illustrate how different participant structures cooccur with differential patterns of teacher and student talk (McCutcheon's point c)); he lists the percentages of teacher and student talk for both small group and full group participant structures (these turn out to be relatively familiar "activities" such as "Reading," "Math Lesson," "Letter practice -drill") in two different teachers' classes. Based on his prior analysis of these teachers' approaches, and microanalysis of some of the lessons, Enright claims that the proportion of teacher talk across activities correlates with (not a statistical test; only two teachers are involved anyway) the differences in the teachers' approaches to turn-taking in participant structures (among other differences). The argument suffers, however, from a failure to recognize point b) above: EVIDENCE. There is a lack of adequate quantitative analysis or illustration of the "microanalysis" of specific participant structures to demonstrate that there are in fact relationships (of a causal or other associative nature) between these variables. For example, only the range of the two teachers' talk is highlighted (and a selected range for one of them as well). They differ little in either full range (proportion of teacher talk is 55.4% to 76.7%, versus 42.7% to 73.2%) or central tendency (medians of 62% and 63.3%, respectively), and the quite similar ranges of student talk are omitted from the discussion. While Enright's full analysis might have the potential of providing significant new insights, without appropriate analysis of observed events, and their use in documenting the occurrence of specific participant structures, the substance of the construct is not validated.

Three Approaches to Validation Using Observation

As I have elaborated on in greater detail elsewhere (Chaudron 1988), several classroom researchers employing observational schemes have attempted validations of their systems in different ways. In a recent review of L1 classroom observation systems, Hoge (1985) showed that many studies demonstrated low validity. He defined three types of validity: *construct validity, criterion-related validity,* and *treatment validity*. In the following, the application of each of these in L2 classroom research will be illustrated.

The most typical method adopted for *construct validation* is to correlate overall scores on some classroom behaviors with separate scores of these behaviors obtained with parallel measures (as in Campbell & Fiske's 1959 classic multitrait-multimethod approach, where multiple traits are assessed each by multiple methods). This procedure in effect substantiates that the constructs involved in the scheme accurately reflect the behaviors they define. Such a procedure, namely correlating teacher ratings of occurrence of various events with low-inference tallying of categories representing those events, was suggested by Ullmann and Geva (1982) as a possible validation of their TALOS system. Otherwise, it has not been widely adopted in recent research (though see Moskowitz' 1976 example of a similar approach, and a critique of it in Chaudron 1988:25).

A form of *criterion-related validation* was performed by Fröhlich, Spada, & Allen (1984) when they attempted to establish a relationship between programmatically defined degrees of communicative language teaching, and the combined values from several independent dimensions of classroom events (on their Communicative Orientation of Language Teaching—COLT—scheme). This commendable effort[3] is unfortunately rare in the L2 research literature. Spada (1987), Allen, Carroll, Burtis & Gaudino (1987), and Lightbown & Spada (1987) have applied this COLT scheme in further efforts to relate observed classroom processes with learning progress in both English and French as a second language (measured by pre- to post-test improvements on various measures). The use of the instrument in these studies resembles that of "treatment" validation (to be described later in this section), with the limitation being that the researchers did not have control over the supposed implementation of programmatic or methodological innovations. Their results

[3] It is flawed in several analytical respects; see Chaudron 1988:27.

have tended more to demonstrate program-or method-internal variability on the observational categories, so that the investigators are led to explore only specific relationships between individual category differences among classrooms and student learning.

While these correlations are a fruitful source of new hypotheses, they constitute neither further validation of the instruments, nor a direct validation of the independent effects of instruction. Spada (1987) and Allen, et al. (1987) are careful to demonstrate, in fact, the extent to which certain of the quantitative analyses derived from the COLT tend to obscure other critical qualitative features of their observed classes (such as the nature of interactive discourse, within a category such as "formal" focus), which interact with the categorial observations. Such findings lead one at first to seek refinement or addition of definitions of certain categories (such as negotiation and concreteness of feedback) that are theoretically or empirically justified as significant to instruction. Further, researchers would prefer to control such important variables more carefully when implementing studies of instructional variables. Both refinement and increased control are part of the continual process of evolution in classroom research referred to in Chaudron (1986a).[4]

Treatment validity refers to the determination of whether observational measures are sensitive to direct intervention on the points being observed. It has too rarely been instituted in L2 classroom research. This approach fits within formative (process-oriented) evaluation procedures, as discussed by Long (1984), where continued observation of classroom processes follows the implementation of new curriculum, teaching approaches, or materials. The lack of such research has of course limited the (internal) validity of many L2 educational comparisons, because the demonstration of delivery of the treatment was neglected (see Long 1984 for further arguments), and only product outcomes were evaluated. Nevertheless, one recent methodology comparison experiment (Bejarano 1987), and one curriculum innovation

[4] Researchers are, however, limited by their lack of responsibility for or involvement in the initial curriculum changes, so that they typically must accept wide program-internal variations as a given. This problem was quite evident in the longitudinal bilingual education program comparison conducted by Ramirez, Yuen, Ramey, & Merino (1986; cf. Chaudron 1988 for some discussion), in which a very detailed (and reliable) analysis of classroom speech act types demonstrated the same sort of intra- program variability found in the COLT-based research (cf. also Nystrom, Stringfield and Miron's 1984 finding that bilingual education program intentions were entirely unrealized, discussed in Chaudron 1988).

project (Rea 1987) illustrate the potential as well as some of the difficulties of such a design. In these studies, the classroom processes intended by the new curriculum or predicted by the experimental methodology were documented using an observation schedule.

In the curriculum development study, a project implementing a task-based academic note-taking course, Rea (1987) proposes a model for curriculum validation that includes (1) checks on the construct validity of the curriculum specifications, (2) criterion-related validity of the intended tasks and materials, and (3) "process-referenced" construct and criterion-related validation of teacher input and learner "uptake" (what learners learn). Without proposing the use of any formal observation scheme, Rea illustrates the observational component of validation by counting the number of student learning tasks (over the entire course) which belonged to different categories relevant to the curriculum goals. These were presumably OBSERVED to have occurred, and not merely intended in the lesson plans.

Here the unit of analysis is not specific classroom interaction behaviors or processes, but the TASKS that are the core of the curriculum (just as the COLT scheme uses "activities" as a base unit). No clear evaluation or criterion is offered to determine whether the observed outcome (an apparent emphasis on the process rather than the product of note-taking) was fully satisfactory. Rea's approach seems to lack a direct demonstration of the relative success or failure of each task, in either a process or product sense, and the evaluation rests at the level of documenting the occurrence of the tasks only. Although Rea's tally appears to show a particular proportion of process and product focus at the task level, without prior expectations for the distribution of these, it is difficult to evaluate the "treatment" validity of these observations.

The argument is in fact circular, if all the researcher has to do to implement treatment change is to add or subtract a task (or other behavior), and then simply count the change when it is implemented. Instead, the treatment changes should be measurable by independent criteria (that is, by means of more specific process and product results WITHIN the observed tasks). Allen, et al. (1987) recognize this when they do a dual analysis of not only the degree of "analytical" and "experiential" qualities of activity units among their observed Core French classes, but also the experiential and analytical nature of processes within those activities.

A study reported by Bejarano (1987), with a much more complete explication in Sharan (1984), was part of a larger curriculum experiment in Israel in 1980-81, in which "cooperative learning" techniques were instituted in both native language literature classes and English as a second language classes. The cooperative learning methods under investigation were two rather different approaches, one a peer tutoring technique, and the other a "Group Investigation" technique (this term used in Sharan 1984, was changed to "Discussion Group" in Bejarano 1987). The research team devoted a half a year to in-service training workshops with teachers in three schools in order to implement these techniques, so that the study's pre-test, validating classroom observations, and post-test were administered in the spring term (March through June).

Although some details are sketchy in the otherwise lengthy report (Sharan 1984), the researchers' effort to validate the treatment delivery through observation is noteworthy. Three independent and trained observers (interrater reliability reported at 85%) employed an adapted 20-item observation schedule (coding social interaction) in each of the experimental and control classes (n=33) at two times about six weeks apart. At each observation, ratings were recorded in three 7-minute intervals spaced throughout the 45-minute periods. As reported in Sharan, Kussel, Sharan & Bejarano (1984b), these observations were checked to determine that at least one-third of the recorded observations in each experimental class (with two classes excepted) conformed to the social organization behaviors expected for those techniques.

In order to appreciate the extent to which these observations were sensitive to the experimental training, however, a complete report should have included the precise categories observed and degree of differences in frequency of observations on those that supposedly discriminated between the three methods groups. For, besides these observations, no other discussion is presented to confirm that these classes in fact differed in just the methodologically prescribed ways and NOT IN OTHER WAYS (nor that they did

not differ in those ways prior to the training program, although this possibility is rather unlikely under the circumstances). In other words, there needs to be a rather exhaustive treatment of the predictable variety of ways in which the classes could differ in terms of social interaction (cf. the theoretical, logical criterion in McCutcheon's argument), in order for there to be confidence in the different treatments as the causal factor. This would not be a very serious concern on the part of the critical reader, if it were not for a rather extended discussion in Sharan, et al. (1984b) explaining that the teachers in the ESL study were extremely resistant (to the point of "rebellion") to the institution of the experimental techniques.[5]

Conclusion

The preceding analysis has been intended to clarify not only the problems and successes in use of L2 classroom observation, but to demonstrate the NECESSITY of validation of these observations, as well as the subsequent need to validate instructional goals and efforts by means of such observations. That is, classroom research is not only of interest to professionals for its own sake, or because it might clarify learning processes, but its use is integral to the eventual success of any research concerned with effectiveness of instruction. The issues of validity that I have raised here are of course not the only sources of error and inadequate interpretation and generalization of research findings;

---

[5] There are in addition a variety of questions as to the relative success claimed by Bejarano (1987) for the experimental treatments over the control classes (Zhang 1988), as measured by differential improvement in target language RECEPTIVE skills. The results are rather complex, in that students of different proficiency levels appeared to improve at different rates depending on the specific treatment received (Sharan, Bejarano, Kussel & Peleg 1984a). The highest proficiency students appeared to benefit most from the experimental treatments, although no statistical interaction effect occurred. Furthermore, with regard to the construct validity of the experimental treatments themselves, George Jacobs and Ted Rodgers (personal communication) have pointed out the weakness of the descriptions of the two (especially the peer tutoring treatment) as representative of "cooperative learning." This matter would bring me into arguments of a more theoretical nature than is my intent in this paper.

I would argue nonetheless, that increased attention to the employment of reliable, validated observation procedures and instruments will lead to substantially greater confidence in the findings of classroom research. Such applications are essential for us to document the course and success of language learning from instruction.

Received October 1, 1988

Author's address for correspondence:

Craig Chaudron
Department of English as a Second Language
1890 East–West Road
University of Hawai'i at Manoa
Honolulu, HI 96822

REFERENCES

Allen, J. P. B., Fröhlich, M., & Spada, N. 1984, The communicative orientation of language teaching: an observation scheme. In J. Handscombe, R. A. Orem, & B. P. Taylor (eds), *On TESOL '83: The Question of Control*. Washington, D. C.: TESOL, 231-252.

Allen, P., Carroll, S., Burtis, J., & Gaudino, V. 1987, The core French observation study. In B. Harley, P. Allen, J. Cummins & M. Swain, *The Development of Bilingual Proficiency: Final Report, Volume II: Classroom Treatment*. Toronto: Ontario Institute for Studies in Education, 56-189.

Allwright, R. L. (ed) 1975, *Working Papers: Language Teaching Classroom Research*. Essex: University of Essex, Department of Language and Linguistics.

Bailey, K. M. 1983, Competitiveness and anxiety in second language learning: looking at and through the diary studies. In H. W. Seliger & M. H. Long (eds), *Classroom Oriented Research in Second Language Acquisition*. Rowley, Mass.: Newbury House, 67-102.

Bejarano, Y. 1987, A cooperative small-group methodology in the language classroom, *TESOL Quarterly*, 21, 483-504.

Brinberg, D. & Kidder, L. H. (eds) 1982, *Forms of Validity in Research*. San Francisco: Jossey-Bass.

Brinberg, D. & McGrath, J. E. 1982, A network of validity concepts within the research process. In Brinberg & Kidder (1982), 5-21.

Brinberg, D. & McGrath, J. E. 1985, *Validity and the Research Process*. Beverly Hills: SAGE Publications.

Campbell, D. T. & Fiske, D. W. 1959, Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin*, 30, 81-105.

Campbell, D. T. & Stanley, J. C. 1972, *Experimental and Quasi-Experimental Designs for Research*. New York: Harcourt Brace Jovanovich.

Chaudron, C. 1986a, The interaction of quantitative and qualitative approaches to research: A view of the second language classroom, *TESOL Quarterly*, 20, 709-717.

Chaudron, C. 1986b, Reliability and validity of categories of classroom discourse analysis. Paper read at the 20th annual TESOL Convention, Anaheim, March 1986.

Chaudron, C. 1988, *Second Language Classrooms: Research on Teaching and Learning*. New York: Cambridge University Press.

Cone, J. D. 1982, Validity of direct observation assessment procedures. In D. P. Hartmann (ed), *Using Observers to Study Behavior*. San Francisco: Jossey-Bass, 67-79.

Edwards, A. D. & Westgate, D. P. G. 1987, *Investigating Classroom Talk*. London: The Falmer Press.

Enright, D. S. 1984, The organization of interaction in elementary classrooms. In J. Handscombe, R. A. Orem, & B. P. Taylor (eds), *On TESOL '83: The Question of Control*. Washington, D. C.: TESOL, 23-38.

Færch, C. 1986, Rules of thumb and other teacher-formulated rules in the foreign language classroom. In G. Kasper (ed), *Language, Teaching and Communication in the Foreign Language Classroom*. Aarhus, Denmark: Aarhus University Press, 125-143.

Fanselow, J. F. 1977, Beyond 'Rashomon' -conceptualizing and describing the teaching act, *TESOL Quarterly*, 11, 17-39.

Folger, J. P., Hewes, D. E., & Poole, M. S. 1984, Coding social interaction. In B. Dervin & M. J. Voigt (eds), *Progress in Communication Sciences, Volume 4*. Norwood, New Jersey: Ablex, 115-161.

Fröhlich, M., Spada, N., & Allen, P. 1985, Differences in the communicative orientation of L2 classrooms, *TESOL Quarterly*, 19, 27-57.

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

Hoge, R. D. 1985, The validity of direct observational measures of pupil classroom behavior, *Review of Educational Research*, 55, 469-483.

LeCompte, M. D. & Goetz, J. P. 1982, Problems of reliability and validity in ethnographic research, *Review of Educational Research*, 52, 31-60.

Lightbown, P. M. & Spada, N. 1987, Learning English in intensive programs in Quebec schools: 1986-87. Unpublished ms., Montreal.

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

Long, M. H. 1984, Process and product in ESL program evaluation, *TESOL Quarterly*, 18, 409-425.

McCutcheon, G. 1981, On the interpretation of classroom observations, *Educational Researcher*, 10, 5-10.

Mitchell, R., Parkinson, B., & Johnstone, R. 1981, *The Foreign Language Classroom: An Observational Study*. Stirling Educational Monographs No. 9. Stirling: Department of Education, University of Stirling, Scotland.

Moskowitz, G. 1976, The classroom interaction of outstanding foreign language teachers, *Foreign Language Annals*, 9, 135-143, 146-157.

Nystrom, N. J., Stringfield, S. C., & Miron, L. F. 1984, Policy implications of teaching behavior in bilingual and ESL classrooms. Paper read at the 18th Annual TESOL Convention, Houston, March 1984.

Page, T. J. & Iwata, B. A. 1986, Interobserver agreement: History, theory, and current methods. In A. Poling & R. W. Fuqua (eds), *Research Methods in Applied Behavioral Analysis*. New York: Plenum, 99-126.

Poole, M. S. & McPhee, R. D. 1985, Methodology in interpersonal communication research. In M. L. Knapp & G. R. Miller (eds), *Handbook of Interpersonal Communication*. Beverly Hills: SAGE Publications, 100-170.

Ramirez, A. 1988, Analyzing speech acts. In J. L. Green & J. O. Harker (eds), *Multiple Persepctive Analysese of Classroom Discourse*. Norwood, New Jersey: Ablex, 135-163.

Ramirez, J. D., Yuen, S. D., Ramey, D. R., Merino, B. 1986, *First Year Report: Longitudinal Study of Immersion Programs for Language Minority Children*. Arlington, Virginia: SRA Technologies.

Rea, P. 1987, Communicative curriculum validation: A task-based approach. In C. N. Candlin & D. F. Murphy (eds), *Language Learning Tasks*. Englewood Cliffs, New Jersey: Prentice-Hall International, 147-165.

Sharan, S. 1984, *Cooperative Learning in the Classroom: Research in Desegregated Schools*. Hillsdale, New Jersey: Erlbaum.

Sharan, S., Bejarano, Y., Kussel, P., & Peleg, R. 1984a, Achievement in English language and in literature, In S. Sharan, 46-72.

Sharan, S., Kussel, P., Sharan, Y., & Bejarano, Y. 1984b, Cooperative learning: Background and implementation of this study, In S. Sharan, 1-45.

Sinclair, J. McH. & Coulthard, M. 1975, *Towards an Analysis of Discourse.* London: Oxford.

Spada, N. M. 1987, Relationships between instructional differences and learning outcomes: A process-product study of communicative language teaching, *Applied Linguistics*, 8, 137-161.

Trueba, H. T., Guthrie, G. P., & Au, K. H-P. (eds) 1981, *Culture and the Bilingual Classroom: Studies in Classroom Ethnography*. Rowley, Mass.: Newbury House.

Ullmann, R. & Geva, E. 1982, Classroom observation in the L2 setting: A dimension of program evaluation. Modern Language Centre, Ontario Institute for Studies in Education (ms.).

van Lier, L. A. W. 1982, *Analyzing interaction in second-language classrooms*. Ph.D. Dissertation, University of Lancaster, Lancaster, England.

van Lier, L. A. W. 1984, Discourse analysis and classroom research: A methodological perspective, *International Journal of the Sociology of Language*, 49, 111-133.

Zhang, S. 1988, Comments on Yael Bejarano's "A cooperative small-group methodology in the language classroom", *TESOL Quarterly*, 22, 2, 347-349.