Center for Second Language Classroom Research

Social Science Research Institute

University of Hawaii

Technical Report No. 8

Reliability and Validity in Second Language Classroom Research

Craig Chaudron, Graham Crookes and Michael H. Long

December, 1988

RELIABILITY AND VALIDITY IN

SECOND LANGUAGE CLASSROOM RESEARCH

Craig Chaudron, Graham Crookes and Michael H. Long

Abstract

The concepts of reliability and validity are defined and discussed with reference to second language classroom research. Examples of reliability and validity assessment are provided. It is suggested that too little attention to these concepts has lessened the meaningfulness of research on second language teaching.

Acknowledgements

We would like to thank Merrill Swain for comments on a portion of this report.

Dec 1988

Technical Reports
of
The Center for Second Language Classroom Research

Social Science Research Institute, University of Hawaii

The opinions expressed in Technical Reports of the Center for Second Language Classroom Research are those of the authors, and do not necessarily reflect the views of the Social Science Research Institute, of the University of Hawaii, or of the funding agency. This report was published by the Social Science Research Institute, University of Hawaii. All rights reserved. No parts of this Report may be reproduced in any form or by any means without the prior written permission of the Director, Social Science Research Institute, University of Hawaii.

Additional copies of this and other Technical Reports may be obtained for $2.00 each, which includes printing and postage, by writing to the Director, Center for Second Language Classroom Research, Social Science Research Institute, c/o Moore Hall 570, 1890 East-West Road, University of Hawaii, Honolulu, Hawaii 96822, U.S.A. Checks should be made out to the "Research Corporation of the University of Hawaii." Prices are subject to increases due to our costs.

1. Michael H. Long, Cindy Brock, Graham Crookes, Carla Deicke, Lynn Potter, and Shuqiang Zhang. 1984. **The effect of teachers' questioning patterns and wait-time on pupil participation in public high school classes in Hawaii for students of limited English proficiency.**

2. Michael H. Long. 1985. **Bibliography of Research on Second Language Classroom Processes and Classroom Second Language Acquisition.**

3. Graham Crookes and Kathryn A. Rulon. 1985. **Incorporation of Corrective Feedback in Native Speaker/Non-native Speaker Conversation.**

4. Graham Crookes. 1986. **Task Classification: A Cross-Disciplinary Review.**

5. Craig Chaudron, Jan Lubin, Yoshi Sasaki, and Tom Grigg. 1986. **An Investigation of Procedures for Evaluating Lecture Listening Comprehension.**

6. Graham Crookes. 1988. **Planning, Monitoring, and Second Language Development: A Review.**

7. Craig Chaudron, Janice Cook, and Lester Loschky. 1988. **Quality of Lecture Notes and Second Language Listening Comprehension.**

Center for Second Language Classroom Research

Information

The Center for Second Language Classroom Research (CSLCR) was established at the University of Hawaii at Manoa in the fall of 1983. Administratively part of the University's Social Science Research Institute (SSRI), it is a joint venture of SSRI and the Department of English as a Second Language (ESL).

SSRI has supported a position for the Center's Director (currently Dr. Karen Watson-Gegeo), who is a faculty member in the Department of ESL, plus administrative and technical support. Supported by internally generated funds and outside contracts and grants, additional CSLCR staff have been drawn from the faculty and students of the Departments of ESL, Linguistics, and Educational Psychology.

The work of the CSLCR includes research, curriculum development and training projects in the general area of second language (SL) education. This includes basic and applied research on SL teaching and learning, on education through the medium of a second language, and on classrooms where second dialects are present (e.g. Hawaiian Creole English). English and other second languages are included in this work. The Director of the Center coordinates all research projects and actively pursues new projects and collaborations with other agencies.

Contents

Reliability and Validity in Second Language Classroom Research

Craig Chaudron, Graham Crookes, and Michael H. Long

Introduction

If second language (SL) teaching is to improve, the characteristics of successful techniques and programs must be disseminated within the SL teaching community. Determining the sources and nature of successful SL teaching requires careful observation and controlled investigation. Research in this field attempts to describe, explain, or understand the instructional processes that result in learning. In every case, the utility of research is related to its internal consistency and quality, and to the meaningfulness of its content to teachers in situations different from the particular instance described. In ordinary terms, it must be *reliable,* and *valid.*

The first stage in a scientific research program (as opposed to a single study) is usually description. If we do not have a good description of something, it is very difficult if not impossible to explain or understand it. Furthermore, as humans have limited abilities to process information, we usually address the problem of describing a complex phenomenon by developing a system for organizing and classifying information. This simplifies the problem by dividing it into conceptually distinct, manageable parts, and it also simplifies the matter of controlling and comparing the information obtained. Therefore, to investigate second language teaching in formal situations (ideally, classrooms) we need concepts and descriptive terms for observable behaviors, interactive events, and types of activities and learning tasks, on which we can build theories of interaction, instruction, and learning.

In this report, we are concerned with the reliability and validity of SL classroom research. With regard to the conceptually simplest kind of research, description, we will merely need to investigate the observational systems employed. In more complex research, where experimenters deliberately intervene in a learning situation to probe more deeply into a phenomenon of interest, we also need to be concerned with the intervention procedures used. Second language classroom research shares most of its research techniques with other domains of educational research, and so we will not need here to reiterate the principles of design and analysis of general interventionist, experimental procedures. Classroom research differs somewhat, however, in its extensive use of measuring instruments based on human observation and classification, as opposed to supposedly more objective (or reliable) instrumentation, such as pencil-and-paper tests and reaction time measures. Although observational instruments are increasingly employed, their use has not in general been the subject of close scrutiny in the methodologically developing field of second language classroom research.

We have organized this report in terms of two prime criteria for scientific investigations: reliability and validity. For present purposes, we will accept the following general definitions of these terms. Reliability is

> the extent to which ... any measuring procedure yields the same results on repeated trials. (Carmines & Zeller, 1979, p. 11)

Validity is the

> degree to which [a] finding is interpreted in a correct way. (Kirk & Miller, 1986, p. 20)

It should be pointed out here, however, that an absolute separation need not be maintained between the two terms, and indeed some authorities (e.g., Brinberg & McGrath, 1985) regard reliability as an aspect of validity. We concur with this view, and will elaborate on Brinberg & McGrath's formulation in the second part of the report.

The report is divided into two main sections, dealing with each of these sub-topics, and a conclusion.

RELIABILITY

Reliability of what?

In discussing reliability in terms of classroom observation, Rowley (1976) points out that although there is a tendency to speak of the reliability of an observational system, this is a rather loose usage. Actually, a given coding system, or **instrument**, is used to obtain an observational **record**, or data. Data may be summarized by way of numerical scores for a given research object, such as number of teacher solicits per lesson, or average length of student utterances, thus forming an observational **measure**. Reliability is not an inherent property of an instrument, but emerges under conditions of use, with regard to the scores derived from the observations. We should also note that the reliability of a set of scores need not be unitary, in that for an instrument of more than minimal complexity, some scores may be reliable and others not.

By considering the stages which must be passed through for an observational record to be obtained, we may arrive at an appreciation of the points which need to be considered for that record to be a reliable one, as well as some indications of the applicability of different measures of reliability. We will first discuss training procedures and conditions of use of the observation system, and then review reliability indices.

Development of an instrument

In developing an observational measuring instrument, that is, one whose use requires raters or observers, consideration has to be given to the complexity of the instrument overall, and the interpretation level of individual codes. By complexity we refer to number of codes and range of behaviors which must be both identified and recorded. Ignoring for the moment other factors, a complex instrument, particularly one requiring real-time coding, is less likely to result in reliable scores (Mash & McElwee, 1974; and cf. Frame, 1979; Kazdin, 1977). Interpretation level refers to descriptions of behavior which range from those called "topographical" (Hawkins, 1982, p. 27) or "low-inference," to those which have been termed "functional" (Hawkins, 1982), or "high-inference." A low-inference code calls for behavior to be described principally in terms of what can be immediately perceived by the observer, largely ignoring the intent of the object observed. For example, a teacher's questions might be tallied as yes/no or *wh*-questions, purely according to their linguistic form. High-inference codes tend to require the observer to determine the intent or function of the observed behavior behind the surface manifestation of the act. Thus, the same teacher's questions might be classified as 'referential' or 'display,' or according to their level of cognitive complexity. Other things being equal, it is more difficult to obtain adequate reliability for scores derived from an instrument which has principally high-inference codes, precisely because of the greater scope for subjective interpretation which they allow. Some instruments may contain both types, especially if behaviors are considered to be hierarchically related.

Training raters

Observer training is always necessary. Most social science which makes use of humans to record data, with the potential for error this implies, requires some indication of the reliability of the data obtained.[1]

The first, and perhaps optional, stage of observer training involves familiarizing observers or raters with the objectives of the study and its setting. At first glance, it might seem obvious that workers in a study should understand its overall intent. However, there has been debate over the possibility of experimenter effects on research (Rosenthal, 1976), and it might be wondered whether in the case where observers are familiar with the hypotheses of the study, this could influence their recording of data. Kazdin (1977) summarizes a number of investigations of expectancy effects on observers' behavior and data recording to the effect that

> expectancies alone are not likely to influence behavioral observations unless some feedback is also provided... Any feedback given to observers should be restricted to the accuracy of their observations, rather than for changes in the client's behavior. (pp. 147-8)

[1] There are some disciplines where, because of the specialized nature of the observation and conditions of observation, direct, quantal indications of data reliability are not provided—non corpus-based linguistics being one obvious example. Anthropology has an alternative means for showing reliability, to whit, the detailed nature of the single-person record obtained and the extended time duration of observations. Even so, it can take 50 years for the unreliability of a particular, prominent data set to be established (Mead, 1928; Freeman, 1983).

After they have received a briefing concerning either the purpose or broad characteristics of the study, and the conditions under which data will be collected and analyzed, raters need to become familiar with the base unit(s) into which behavior is to be separated, e.g., turns, utterances, T-units, and communication units. It is desirable for written definitions and unambiguous examples of these units to be made available to raters. Definitions should be discussed until they are clearly understood. At a second level of analysis, the categories to which units can be assigned must be defined and raters must be trained in the application of these codes, such as 'general and personal solicits,' 'inform,' 'praise,' and 'correct.' (See e.g., Hawkins, 1982, for examples of definitions.)

The verbal definitions of the categories can affect reliability. Although a high-inference code may cause problems for raters, if it can be very clearly defined, it is less likely to lead to observer disagreement. Contrariwise, segmenting or coding into even low-inference units of behavior which are not clearly specified may lead to difficulties.

If behavior descriptions, or codes, can be developed in collaboration with observers, there is less likelihood of final scores being unreliable. When observers are required to code behavior in units which are congruent with how they perceive the research situation, there is less chance of misunderstandings resulting in lack of reliability. This option is probably not available unless raters are, for example, students or professionals trained in areas related to the topic of investigation.

In the next stage of the training process, sample materials (e.g., audiotapes, videotapes, or transcripts) should be segmented and coded, and agreement between trainees and a criterion rater[2] measured (see Chaudron, Cook & Loschky, 1988, for a detailed description of this process in a SL research context; also Chaudron, 1988, p. 24, n. 2). Disagreements should be discussed, so that eventually raters will reach a high level of agreement with the criterion rater. This inter-rater agreement would be set at some conventionally acceptable level—90% is the most commonly used level, or *kappa* above 0.6 (see below). Slightly lower levels (e.g., 85%) might be acceptable if the coding system involves many high inference categories. Failure to get close to these levels of agreement following sustained training efforts should be taken as an indication that either the coding system is ill-conceived, or the objects or behavior to be rated do not correspond to the system (see, e.g., Crookes, 1986). Coders in SL research are often working with the developers of the rating system, so if the source of the trouble can be identified it may be possible to introduce an immediate modification of, for example, the verbal definition of a given coding category, which will increase inter-rater agreement. Alternatively, where an isolated piece of behavior or language is the cause of disagreement, raters may devise a convention that it be classified in a particular fashion. A few cases of this sort should not be unexpected (de Beaugrande & Dressler, 1981; Frick

[2] A criterion rater is someone accepted as expert in the system concerned, often the individual who devised the instrument or who has ben previously trained and attained acceptably high reliability in its use.

& Semmel, 1978). However, if more than a handful of items seem to require such treatment this would again be evidence that the overall system needs revision, or that a different system should be applied to the corpus at hand. Only following this sort of coder training and instrument development is it legitimate to begin to code data for research purposes.[3]

Use of an instrument

Conditions of use can also limit the reliability of an observational/discourse analysis (DA) instrument. Trivially, the usual conditions for reliable human judgment apply to the use of observational/DA systems (adequately comfortable environment, appropriate duration of observation, sufficiently motivated or remunerated workers). More substantively, the nature of the instrument itself can ultimately affect the reliability of scores obtained. A preliminary consideration is whether or not an observer is required to enter the classroom, or at least be in a position to be perceived by the individuals being observed. If observation results in "reactive effects" (see, e.g., Kazdin, 1982) this may (*ceteris paribus*) lower reliability by way of increasing the instability of behavior. (It will also affect validity.) In recent years, classroom observers have increasingly made use of videotaping. This enables subsequent detailed analysis of the record under optimum conditions, though it may result in some loss of information which could have been gathered by the more sensitive human observer. On the other hand, the active nature of a human observer taking notes of particular behavior may be more intrusive than video equipment, which (if not part of the observation classroom) is often placed in a corner, and having been set going, needs little further attention. We are not aware of any attempt to assess the effects of such observation types on reliability of the records collected in SL classroom research (but see Renne, Dowrick & Wasek, 1983, for some non-educational investigations).

Continuing use of an observational instrument may change patterns in the way observers use it. Raters should agree not only with the criterion rater and with each other, but also with themselves over time, the latter being known as **intra-rater** reliability. Once trained, there is no guarantee that raters will remain faithful to the original interpretations they have been trained to make. Extended studies making use of observational systems should conduct periodic checks of rater drift, and if necessary conduct retraining sessions. In fact, however, almost no SL studies have reported such controls. This may be because SL investigations rarely have the duration of the large-scale mainstream educational studies on which such procedures were developed (see e.g., the various studies reported in Medley & Mitzel, 1963). This state of affairs may be expected to change with increasing demands for accountability in SL education, too,

[3] We mention only briefly here that, following the attainment of high reliability in training, if multiple coders/raters continue to be used to code the same behaviors, the researcher will be faced with decisions as to how to make the separate ratings or codes converge, combine, or average into one set of values for the study. This problem is theoretical and depends on the nature of the instrument and the research purposes—it is not strictly a question of reliability.

however, whereupon intra-rater reliability checks will presumably become more common.

Reliability indices

A variety of procedures are available for calculating measures of reliability and inter/intra-rater agreement. There are some conceptual problems associated with them, however. For instance, Frick & Semmel (1978) comment that there has been a mistaken tendency to accept observer agreement coefficients as equivalent to reliability coefficients (see also Mitchell, 1979).

> Observer agreement is a primary, but not the most crucial, issue in the interpretation of results of observational studies. Most important are the reliabilities of the measures of the **subjects** of observation, that is, the extent to which differences in teachers/learners and the contexts in which they behave are dependably discriminated by the observational measure. (p. 158)

Following standard measurement theory, this type of reliability is defined as the ratio of the variance of true scores of subjects to the variance of observed scores. It is equivalent to the intraclass correlation coefficient (Ebel, 1951; Hoyt, 1941), which may be obtained as follows (Rowley, 1976, pp. 53-4):

> Suppose that $n>1$ visits are made to each of t teachers, and on each visit, an estimate is made of some characteristic X of that teacher. The visits are regarded as equivalent to one another...Application of the standard one-way analysis of variance to these data, with Teachers being the only factor, and visits treated as replications...yield[s]...the mean square associated with the factor "Teachers" [MS_t]... and the ... mean square within teachers [MS_w]. The reliability of a score obtained from a single visit [is]
>
> $$(MS_t - MS_w)/(MS_t + (n\text{-}1)MS_w).$$

This may be more intuitively comprehensible as

> the extent that the average difference between two measurements independently obtained in the same classroom is smaller than the average difference between two measurements obtained in different classrooms. (Medley & Mitzel, 1963, p. 250)

Some factors will diminish the size of the reliability coefficient obtained in this fashion. Instability of the observational source (teacher, pupil, etc.) is the greatest source of unreliability according to Medley & Mitzel (1958, 1963), who accordingly suggest obtaining at least twelve observations to compensate for this. Unreliability will also appear when there is little variation in the behavior being observed, so that systematic variation (that which is intended to be measured) becomes small in relation to random error.[4] Finally, insofar as it is a component of the total error variance, inter-rater

[4] Some technical problems also exist in the caluculation and test of intraclass correlations (Lahey, Downey & Saal, 1983).

reliability places a top limit on the reliability of the scores obtained on a given instrument as measured above.

Despite the theoretical precedence of the intraclass correlation coefficient as a measure of reliability, Frick & Semmel (1978) question its utility. They point out that intraclass correlation coefficients are not useful for measuring observer agreement before or during a study, since using them to assess observer agreement depends on having an estimate of the true variance of the research object(s). This would presume the prior existence of sets of studies of the same phenomenon—something which is very rare in SL classroom research. Medley & Mitzel (1958, 1963, p. 309) assume that a reliability study will be done on an observational instrument before any proposed study using it is actually carried out. While this may well be desirable, considering the amount of time and effort needed to develop a SL classroom observational system, and the relatively limited availability of funds, trained raters, and access to classrooms, it will be a very rare classroom research project which has this luxury.

Accordingly, while recognizing the conceptual primacy of intraclass correlation measures of overall reliability, we will in the remainder of this section follow general practice in SL classroom research/DA and focus on purely observer-based indices of reliability.

Inter-rater reliability measures depend on the nature of the categories used, the events being observed, and the units used for segmentation. An initial distinction has to be made as to whether individuals must separate the stream of behavior into units, which are later to be categorized, or whether they are simply to note the occurrence of specific behavioral acts. For example, is the raters' first task to segment teacher speech into utterances (or turns), the utterances (production unit) later to be classified according to their function, or is it at once to record the occurrence of such items as questions (of various kinds), corrections, and so on? Reliability may be affected by the rapidity with which such decisions must be made (see discussion of some of these issues in Long, 1980).

Furthermore, observers may be coding only for behaviors occurring within specified time periods. A time-sampling system can, however, lead to complications concerning both reliability and validity. Particularly in real time coding systems, where there are a large number of behaviors to be recorded, it is obviously impossible for observers to both observe continuously and record data. So they may be required to observe for, say, 15 seconds, and then record for the same time period. Depending on the actual frequency of the behavior to be observed, and the observe/record periods, time-sampling can result in quite inaccurate estimates of the overall frequency of occurrence of the behavior of interest. Although we suspect that this point may not yet be familiar to many SL researchers, the danger is not too serious, partly because we suspect real-time observation is being less used in favor of video- or audiotaping and subsequent analysis, and also because a summary of the problem and solutions has been laid out by Rojahn & Kanoy (1985), to which we refer the interested reader.

In the first of the two cases mentioned above, where individuals have first to segment the behavioral stream prior to categorizing the units thus identified, inter-rater agreement indices will be needed for both phases. An index for the first phase (called "unitizing reliability" by Folger, Hewes & Poole, 1984, p. 118) is provided by Guetzkow's (1950) *U*. This is given by

$$(O_1 - O_2) / (O_1 + O_2)$$

where O_1 is 'number of units identified by observer 1', and O_2 is 'number of units identified by observer 2'. Use of this index has largely been confined to research in the communication sciences, though it would apply equally elsewhere. Guetzkow's index is actually a measure of disagreement, so small scores (below 0.1 has been suggested) are desirable. As with most inter-rater reliability indices, it has some limitations (see Folger, et al., 1984, for further details). A formula used in early reliability studies (Osgood, Suci & Tannenbaum, 1957) and some recent work (Kreckel, 1981, p. 108) is

$$V = \frac{2 N_{ij}}{N_i + N_j}$$

where N_{ij} is the number of units segmented the same by rater i and by rater j, and N_i and N_j are the number of units identified by each rater. A simple percentage agreement score can also be obtained, from the following widely used variation on the above:

$$N_a / (N_a + N_d) \times 100\%$$

where N_a is 'number of agreements' and N_d is 'number of disagreements'.

A great deal of attention has been devoted to the development of measures for the second, 'interpretation' phase. For many studies, this is the only reliability index that will be obtained. In one summary paper, Berk (1979) deals with twenty-two different such indices, and many more such reviews are available (see Hartmann, 1982, for a brief review of reviews). Obviously, it is not possible even to list the different statistics here. The most conceptually transparent, simplest, and probably most common measure found in second language research using coding or observational systems is some version of simple agreement expressed as a percentage form (like that mentioned immediately above).

At least four variants on this are possible, as Page & Iwata (1986) make clear. *Total agreement* applies when two observers each separately record the number of instances of a particular behavior they observe in a given time period. Dividing the smaller of these two figures by the greater and multiplying by 100 gives a percentage value with a maximum of 100%. In fact, of course, the two observers may not actually have been agreeing on what they saw. If observer A sees something she believes to be an instance of the desired behavior during minute 3 and during minute 6 of a ten-minute period, and observer B sees something supposedly indicative of the behavior of interest at minutes 5 and 9, both will have recorded two instances of the behavior in the time period, and under this simple system will have obtained perfect agreement, despite actually having exhibited inconsistency of observation.

If it is possible to break down the overall period to a minute-by-minute record, then a more sensitive measure for this series of observations can be obtained. During minutes 1 and 2, both observers were in agreement. At minute 3, they were in disagreement, one seeing an instance of the target behavior, the other, not. Proceeding similarly through each minute of observation, the situation may be summarized using the simple formula for agreements given above, where 'number of agreements' is replaced by 'number of agreements that behavior occurred in time period t.' There were 6 agreements and 4 disagreements, which gives *interval agreement* of 60%.

Note, however, that if 'number of agreements' is replaced by 'number of agreements concerning occurrences of behavior,' the observers never agreed that an instance of behavior occurred in the same time period. A percentage measure of occurrence agreement will show 0% in this case. The observers' agreements only concerned nonoccurrences of behavior. If 'number of agreements' is replaced by 'number of agreements concerning nonoccurrences of behavior,' there were again 6 agreements concerning nonoccurrence, and 4 disagreements, again giving a figure of 60%. (See Page & Iwata, 1986.)

Measures of this sort are influenced by the possibilities for chance agreement. As Page & Iwata (1986, p. 112) state,

> two observers randomly scoring data sheets can be expected to agree some percent of the time due to chance alone...When the behavior occurs at a very low rate, chance agreement increases dramatically due to the large number of potential agreements on nonoccurrence, and when the behavior occurs frequently, chance agreement increases due to the potential agreement on occurrence.

Where there are a number of codes for which instances of behavior are being tallied, pair-wise agreement can be calculated via a simple Pearson product-moment correlation coefficient. For example, for this data set:

Behavior type	A	B	C	D	E
Observer 1 # of items	6	6	5	8	6
Observer 2 # of items	4	5	3	8	6

a correlation coefficient of 0.95 is obtained. This can be used as a measure of inter-rater agreement. But consider this set:

Behavior type	A	B	C	D	E
Observer 1 # of items	3	4	2	7	5
Observer 2 # of items	4	5	3	8	6

Here, the two observers are operating rather well, though they do not appear to be in total agreement, as there is a discrepancy of 1 unit of behavior between them on each code. Unfortunately, such a difference will not show up at all if a correlation coefficient is used as a measure of inter-rater agreement, since these two sets of figures are perfectly correlated ($r_{12} = 1.0$).

The SL research community is starting to be aware of the fact that such measures allow chance agreement to enter in, or indicate perfect agreement when it does not actually exist, and thus do not actually provide a particularly good estimate of raters' abilities. Cohen's *kappa* (Cohen, 1960), one of several available statistics which compensate for chance, (see e.g., Brennan & Prediger, 1981) is increasingly being used in educational research. Because of this, we will discuss its use in some detail.

Kappa is a statistic which indicates what proportion of the agreements not expected by chance occur, after removing from the observed agreement figure those agreements expected by chance. That is to say, if all the agreements not expected by chance occur, *kappa* will equal one. If only some of the agreements not expected by chance occur, *kappa* will be less than one. If none of the agreements not expected by chance occur (that is to say, if all of the agreements which do occur are those expected by chance), then *kappa* equals zero. Actually, *kappa* can be less than zero, when fewer than chance agreements occur, but this is not of great interest.

To restate this matter in figures (following Conger, 1985, p. 863), let us take the following example:

		Observer 2		
	Category	A	B	
	A	320	190	510
Observer 1	B	210	280	490
		530	470	1000

Two observers have classified 1000 items as either As or Bs. By the rules of probability, from the marginal totals we can calculate the probability of the observers accidentally agreeing that a given item is an A. This is the probability of Observer 1 identifying any item as an A (0.51) multiplied by the probability of Observer 2 identifying any item as an A (0.53), which is a probability of 0.2703. That is to say, of the 320 cases where the observers agreed that an item was in fact an A, about 270 of those instances might have occurred by accident, given the way the two observers were performing overall.

A similar procedure can be performed for each cell (and the table can be larger than 2 x 2, and indeed more than two observers can be used—see Fleiss, 1971). Completing the procedure for the whole of this table gives us two figures, one for observed agreement overall

$$P_o = 0.32 + 0.28 = 0.60$$

and one for chance agreement overall

$$P_c = (0.51 \times 0.53) + (0.49 \times 0.47) = 0.5006.$$

We are interested in non-chance agreement. In general, the probability of an event not occurring is one minus the probability that it does occur. So the probability of non-chance agreement in this case is 1 - 0.5006, = 0.4994. That is to say, out of the 1000 decisions the table indicates, about 499 agreements might not have occurred by chance. Now, 600 agreements did in fact occur, and expected chance agreements come to 501 (rounding off), so 99 more agreements than might have been expected by chance occurred. Thus, the agreements above and beyond those expected by chance, as a proportion of the total number of non-chance agreements, come to 99 / 499 ≈ 0.19. To summarize in an equation:

$$\textit{kappa} = \frac{P_o - P_c}{1 - P_c},$$

and substituting in the figures obtained above,

$$= \frac{0.60 - 0.5006}{1.00 - 0.5006}$$

so that *kappa* in this case equals 0.19. Obviously, calculating *kappa* is a little more complex than calculating percentage agreement, and the result is a figure which is perhaps less intuitively interpretable, though one which is definitely more desirable in its avoidance of chance agreement. A useful compromise may be to report both percentage agreement and *kappa*, as in Crookes & Rulon (1985). Despite its problems, percentage agreement has the advantage that it is easily interpreted, and reporting *kappa* provides a more careful check.

It is customary in social science research to report the significance of summary statistics (despite criticisms of this for at least 30 years, see, e.g, Rozeboom, 1960). An observed value of *kappa*, may also be tested against the hypothesis that *kappa* is zero (see Fleiss, Cohen & Everitt, 1969; Fleiss, 1971). However, that *kappa* (or any other appropriate measure) is significantly different from zero is not usually informative, since as with Pearson's *r*, even numerically quite small values of *kappa* can be statistically significantly different from 0.[5] As mentioned above, conventional levels of agreement are preferable, and a rule of thumb statement by Hartmann that *kappa* should be larger than 0.6 has been used in some SL research reports (Crookes, 1986; Gelfand & Hartmann, 1975; Hartmann, 1977).

Reliability in second language classroom research

As summarized in Chaudron (1988, p. 25-26), few second language classroom researchers have endeavored to confirm the reliability of their observations or analyses. A few well-known instruments and studies exist, and their developers made efforts to train raters in their use. However, follow-up reliability assessment to check against observer drift has rarely been applied. Moreover, most current instruments need to be used by observers other than those who developed them, in order to determine their inter-rater reliability over time and across contexts.

Summary

We have outlined in some detail here the nature of and procedures for conducting reliability assessment in SL classroom observation and discourse analysis because reliability is underused, and perhaps less well understood, among researchers in this domain than it is in other areas of SL research, such as testing and experimental research. Our intention is not to be critical of past efforts or failures to ensure reliability, for we are fully aware of the labor-intensive nature of the development of measures and the training of observers/analysts to make their use reliable. We prefer simply to urge that future research efforts take greater stock of reliability in their design and implementation, so that greater confidence can be had in their results.

VALIDITY

In the second part of this report, we explore first the nature of validity in classroom research, outlining in particular the place of observation in the validation of second

[5] In addition, traditional significance tests of such measures of agreement have been criticized on the grounds that the data on which the measures are based almost always exhibit autocorrelation, which is a violation of assumptions underlying use of the tests (see Faraone & Dorfman, 1988).

language classroom research with respect to different methodological orientations. We will then illustrate three different approaches to observational validation of instructional research in a critique of recent SL studies. We have two main concerns: how observational analyses of classroom interaction can be validated, and how claims about instructional variables (such as the effectiveness of programs, teaching methods, syllabus changes, materials, rule presentations, and so on) depend on the application of validated observational analyses.

The concept of validity

There are many current discussions of the role of validity in the research process (Campbell & Stanley, 1972; Brinberg & Kidder, 1982; Judd & Kenny, 1982; Brinberg & McGrath, 1982, 1985; Cone, 1982; LeCompte & Goetz, 1982; Folger, Hewes & Poole, 1984; Hoge, 1985; Kliess & Bloomquist, 1985; and Poole & McPhee, 1985). In most of these, the various aspects of validity refer in essence to the determination of the 'truth' of an analysis or theory. We find that the recent formulation of validity by Brinberg & McGrath (1985) provides the most comprehensive framework within which to interpret the other views.

Most discussions of validity follow the general principles developed by Campbell and his colleagues in the past twenty-five years (see especially Campbell & Fiske, 1959; Campbell & Stanley, 1972—a new edition of their 1963 work; and Cook & Campbell, 1979). 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. Cook & Campbell (1979) then refined these two notions by splitting off 'statistical conclusion validity' as a special form of internal validity referring to the correctness of the conclusions from the statistical results, and 'construct validity' as a special form of external validity referring to the relationship between the observed variables and the conceptual independent and dependent variables (see also Rosenthal, 1982, and Kliess and Bloomquist, 1985).

Brinberg & McGrath (1985) have proposed a heuristic model of the research process that encompasses the many varieties of validity in a Validity Network Schema (VNS). We reproduce their Table 1.1 here as Table 1, as the point of departure for our discussion of validity in observation and classroom research. Brinberg & McGrath identify three general domains of interest in research endeavors (the conceptual, methodological, and substantive), and three stages of the research process (see the listing of central tasks for each stage in Table 1); validity has a quite different meaning depending on which stage is involved, and which domain or combination of domains are emphasized within a given stage. Thus, the first stage of the research process is one in which basic elements, relations, and systems are evolved and clarified; validity in this stage is a matter of the 'value' of the elements, relations, etc., with regard to the different criteria appropriate to each domain. This is closest to the notions dealt with in 'construct' validity, although Brinberg & McGrath's formulation is much more complex, due to the differentiation of criteria for each domain.

Brinberg & McGrath propose that in the second stage of the research process, the three research paths ("experimental," "theoretical," and "empirical")[6] each derive from emphasis on two domains in the first two steps of development, leading to either a study design, a set of hypotheses, or a set of observations (*Step* 2). This step is followed by application in the third domain (*Step* 3). Figure 1 illustrates this conceptualization, showing for example, that the combination of conceptual and methodological domains leads to the experimental path and a study design to be implemented in the substantive domain, while the methodological and substantive domains combine in the empirical path, to be interpreted in the conceptual domain. Each of these paths involves variables, in which the 'correspondence' between specific elements and relations result in various types of validation. For example, the variety of aspects referred to generally as 'internal validity' (e.g., maturation, history, instrumentation, selection) are viewed by Brinberg & McGrath as forms of 'logical correspondence' between the concepts studied and the methods used to study them, while 'statistical conclusion validity' is a form of 'empirical correspondence.' Thus, each path toward a set of empirical findings in stage two involves the matching or fitting of elements, relations, and systems of one domain with those of another; correspondence validity is therefore the determination that there is in fact an appropriate matching. So the threat to validity occurring when subjects in an observed group experience some variable that is not controlled or measured ('history') is a failure of appropriate matching between the conceptually significant factors and the substantive ones occurring in the study.

[6] We recognize that Brinberg & McGrath's adoption of such commonly used terms may be confusing, but ask the reader to suspend temporarily any standard interpretation of them in favor of their intended technical use as labels for the path types.

Table 1

The VNS System: Validities and Stages of the Research Process

Stage One: Validity as Value

(Central tasks of stage one: Identification, development; and clarification of elements, relations, and embedding systems, for each of the three domains.)

Domain	*Criteria for Evaluating Elements, Relations*
Conceptual (C)	Parsimony, internal consistency, subsumptive power, testability, etc.
Methodological (M)	Efficiency, power, unbiasedness, explicitness, reproducibility, etc.
Substantive (S)	System effectiveness, cost/benefit, feasibility, etc.

Stage Two: Validity as Correspondence

(Central tasks of stage two: selection, combination, and use of elements and relations from all three domains to produce a set of empirical findings.)

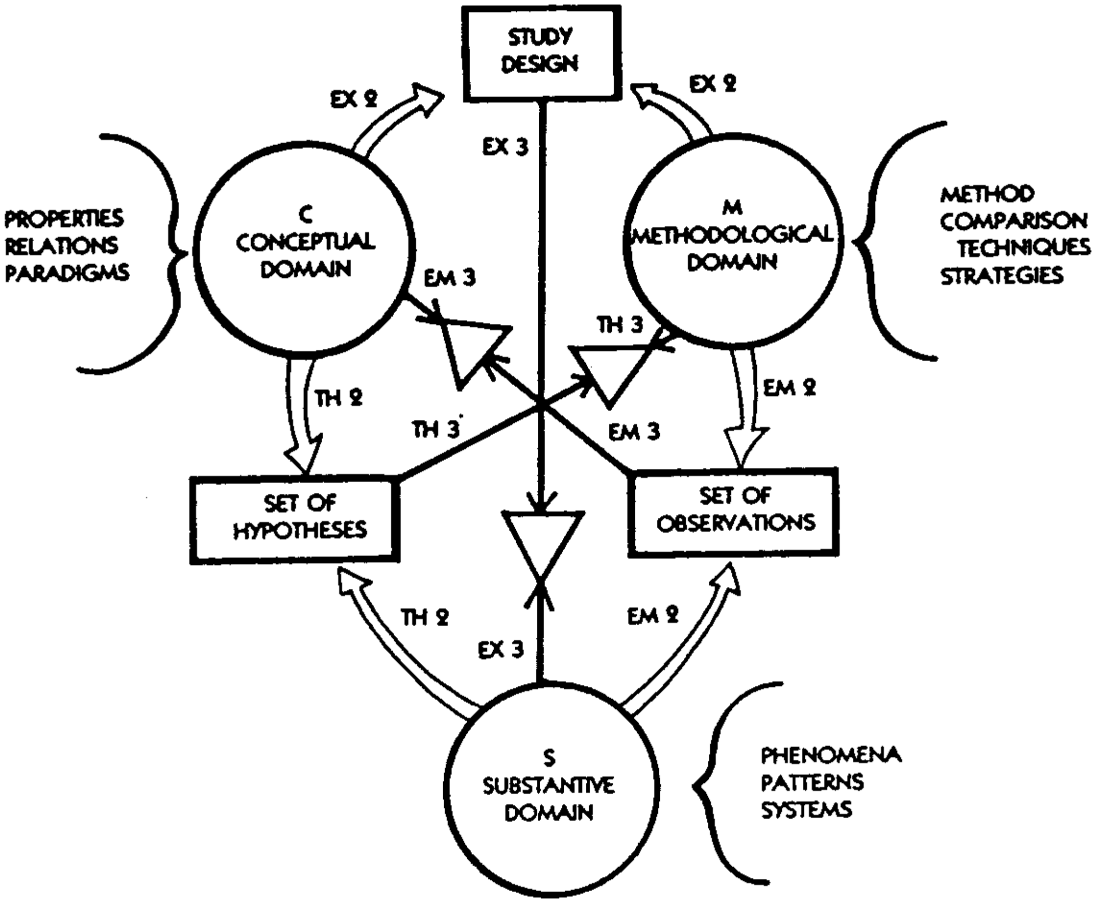
Paths	*Step 2*	*Step 3*	*Product*
Experimental	Study design	Implementation ↘	A set of empirical findings
Theoretical	Set of hypotheses	Test of hypotheses →	
Empirical	Set of observations	Interpretation ↗	

Stage Three: Validity as Robustness

(Central tasks of stage three: verification, extension; and delineation of particular stage two findings.)

Replication:	Are the (stage two) findings reproduced when all facets of C, M, and S are kept the same?
Convergence Analysis:	Over what range (of values of all facets of C, M, and S) do the (stage two) findings hold?
Boundary Search:	Beyond what range (of values of all facets of C, M, and S) do the (stage two) findings fail to hold?

(From Brinberg & McGrath, 1985, p. 23)



EXPERIMENTAL PATH: Building a design, and implementing it by using it on a set of substantive events.

THEORETICAL PATH: Building a set of hypotheses, and testing them by evaluating them with an appropriate set of methods.

EMPIRICAL PATH: Building a set of observations, and explaining them by construing them in terms of a set of meaningful concepts.

Figure 1: The VNS System: Domains, Levels, and Paths
(From Brinberg & McGrath, 1985, p. 22)

Finally, the third stage, involving the extension of findings in research, is on the whole what has been understood as 'external validity.' This brief outline of the VNS does not do justice to its complexity, but in referring to these principles while exemplifying efforts to validate second language classroom research, we hope to clarify it and justify its usefulness.

Validation of observation

Observational procedures and concepts are like any measurement systems (or tests), in that they are subject to the form of validation known as 'instrument validity,' usually determined by means of systematic procedures such as content, construct, and predictive validity (we use these terms in their accepted sense from psychometric research, as in Nunnally, 1978; Thorndike, 1982); part of these procedures are of course the reliability assessments described in the first part of this report.

With reference to the VNS framework, development and use of observational instruments or analyses would be primarily a methodological endeavor in stage two, but not exclusively so, in that conceptually driven questions related to the theoretical nature of interaction, or substantive issues regarding efficiency in classroom grouping may be the underlying motives for developing the instrument. This development is in fact a multi-stage process, in which validation of the instrument and analytical categories can occur at each stage, as suggested above. The first stage would require analysis of the logical relations, comprehensiveness, importance, efficiency, etc. of the categories adopted, (i.e., their 'value'). In the second stage, observational instruments or analyses would be most typically employed as adjunct methods for other research goals, most specifically in ensuring and documenting that relevant treatments, variables, and processes in fact occur (an assessment of 'correspondence' validity). Furthermore, in the third stage, the evaluation of such observational descriptions as constructs relevant to the research questions can only be completed if the observations and summary findings of the study are shown to hold generally in different contexts and at later times ('robustness' validity). Such validation is accomplished through rigorous application of sampling procedures and design principles, and in replications: the same or similar observational analyses must be applied to new situations and populations.

In the following discussion, these validity types will be exemplified and compared to recent related formulations in the literature on classroom research. We will first explore two broad methodological orientations taken by language classroom researchers, and illustrate the criteria by which validity must be evaluated. Then, we will relate three commonly referred to types of validity to Brinberg & McGrath's system, while further illustrating with recent second language classroom research.

Methodological orientations

Researchers can describe classroom events according to various theoretical perspectives, which result in different methodological orientations to observational analysis. In Brinberg & McGrath's terms, these perspectives may derive from unique domains (conceptual, methodological, or substantive), or from combinations of them. As methodological orientations, they would also be primarily oriented toward the development of research in a particular path within stage two. The common orientations of classroom researchers 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 further discussion in Long, 1980; Allwright, 1988). The principal distinctions among these involve the degree to which an exhaustive and structured set of categories of behavior are used to describe the interaction, and whether analyses are theoretically motivated, usually before a study begins, or theory is allowed to arise naturally from the data themselves. Studies vary from descriptive approaches with apparently no presupposed categories, to use of highly complex systems of predetermined categories and dimensions of analysis.

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's COLT, 1984; and others related to these). Interpretive and ethnographic perspectives, on the other hand, 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, 1988; and Bailey, 1983). As is clear from the general educational and psychological literature on validity cited above, and as argued in the first part of this report, such descriptions, like any measurement instruments, must be evaluated for their reliability first, and then for other forms of validity. This is so regardless of the theoretical perspective taken, as LeCompte & Goetz (1982) and McCutcheon (1981) demonstrate quite clearly in discussing validity in ethnographic research. We will consider McCutcheon's general argument.

While McCutcheon (1981) was especially concerned with qualitative research (as in ethnography), her position applies equally well to the quantitative researcher who might, for example, adopt the set of descriptors in Sinclair and Coulthard's (1975) type of hierarchical discourse analytical system (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 claims 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, predictive value, etc.). Furthermore, in assuming an audience for their interpretations, researchers also necessarily expect intersubjective (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 criteria for evaluating interpretations are commonsense expressions of several of the distinct types of validation outlined in Brinberg & McGrath's VNS. The *logic* criterion coincides with their conceptual domain criterion for value, the *evidence* and *consistency* criteria are different aspects of correspondence validity, and the *significance* basis is partly a matter of conceptual and methodological value, partly of the robustness criterion.

We can exemplify these principles using Enright's (1984) predominantly ethnographic study of second language classrooms. He uses the concept 'participant structures,' a 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.

McCutcheon's criteria a) through d) all point to ways in which Enright's analysis needs validation. First, there must be an internal *logic* that systematically interrelates the different participant structures.[7] Demonstrating this internal logic is essentially equivalent to the procedure of verifying that items on a test represent the skills or knowledge individuals are being tested on, commonly referred to as construct validity.

Bypassing *evidence* for the moment, and considering McCutcheon's criterion of *consistency*, or correspondence, it should be clear that observations of participant structures must associate with other behaviors related to them (commonly a matter of concurrent validity), which Enright attempts to demonstrate (see the next paragraph). Further, with regard to the *significance* of this analysis, it must prove to have some bearing on further findings with these or similar teachers and contexts (generalizability or robustness), 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 (the matter of consistency): he lists the percentages of teacher and student talk for both small group and full group participant structures (these turn out to be realized as relatively familiar activities such

[7] However, it should be recalled that 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 agreement: the changes in dimensions of actors, topics, etc. that together constitute different participant structures must be identifiable and recognized by independent observers (intersubjectively).

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); this is clearly a matter of correspondence. The argument suffers, however, from a failure to recognize the need for the further measure of validity mentioned above: *evidence*. There is a lack of adequate quantitative analysis (cf. statistical conclusion validity) 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 differences between specific participant structures, the substance of the construct itself is not validated.

Three approaches to validation using observation

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 the three types of validation in L2 classroom research and their relationship to the VNS 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 behaviors involved in the scheme accurately reflect the constructs they instantiate. Although this method derives from 'value'-oriented concerns for validity, the procedures employed follow one of the research paths in stage two of the VNS system, specifically the theoretical path establishing the correspondence between the substantive domain included in the observation system, and the conceptual domain that specifies particular relationships between the observational categories.

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, p. 25).

A form of *criterion-related validation* was performed by Fröhlich, Spada, & Allen (1984). In the VNS system, this is again a correspondence validity issue, but with the additional aim of the stage three 'convergence analysis' to determine the robustness of the observational system. Fröhlich, et al. (1984) 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 is unfortunately rare in the L2 research literature. Spada (1987), Allen, Carroll, Burtis & Gaudino (1987), and Lightbown & Spada (1987) have applied the 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 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, b).[8]

Treatment validity refers to the determination of whether observational measures are sensitive to direct intervention on the points being observed. In the VNS system,

[8] 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; see 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).

this is more of a substantive issue, focused however on the robustness of the observational system. Tests of such validity have too rarely been instituted in L2 classroom research. This approach fits within formative evaluation procedures, as discussed in 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 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 (inter-rater reliability reported at 85%) employed a 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 for the evaluator 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 conceptual, 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.[9]

[9] 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

Conclusion

The preceding analysis has been intended to clarify not only the procedures, problems, and successes in reliability assessment and validation of L2 classroom observation, but to demonstrate the *necessity* of applying these principles in observational research, as well as the subsequent need to validate instructional goals and efforts by means of such observations. That is, classroom research is not simply 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 the effectiveness of instruction. The issues of reliability and validity that we have raised here are of course not the only sources of error and inadequate interpretation and generalization of research findings; Brinberg & McGrath (1985, especially Chapter 2) have in fact argued that an inevitable result of attempts to maximize validity of one sort is that another sort of validity is diminished. This leads to the proposal that multiple and converging research efforts are necessary in order to fully explore conceptual, methodological, or substantive domains of research.

We 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.

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 us into arguments of a more theoretical nature than is our intent in this report.

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*, (pp. 231-252). Washington, D. C.: TESOL.

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*, (pp. 56-189). Toronto: Ontario Institute for Studies in Education.

Allwright, R. L. (Ed). (1975). *Working papers: language teaching classroom research*. Essex: University of Essex, Department of Language and Linguistics.

Allwright, D. (1988). *Observation in the language classroom*. London: Longman.

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*, (pp. 67-102). Rowley, Mass.: Newbury House.

Beaugrande, R. de & Dressler, W. (1981). *Introduction to text linguistics*. New York: Longman.

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

Berk, R. A. (1979). Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency 83*, 460-472.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient *kappa*: some uses, misuses, and alternatives. *Educational and Psychological Measurement 41*, 687-699.

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 D. Brinberg & L. H. Kidder (Eds.), *Forms of validity in research,* (pp. 5-21). San Francisco: Jossey-Bass.

Brinberg, D. & McGrath, J. E. (1985). *Validity and the research process.* Beverly Hills, CA: SAGE.

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.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: SAGE.

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.

Chaudron, C., Cook, J. & Loschky, L. (1988). *Quality of lecture notes and second language listening comprehension.* Technical Report no. 7, Center for Second Language Classroom Research, Social Science Research Institute, University of Hawaii.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20,* 37-46.

Cone, J. D. (1982). Validity of direct observation assessment procedures. In D. P. Hartmann (Ed), *Using observers to study behavior,* (pp. 67-79). San Francisco: Jossey-Bass.

Conger, A. J. (1985). *Kappa* reliabilities for continuous behaviors and events. *Educational and Psychological Measurement 45,* 861-868.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Chicago: Rand McNally.

Crookes, G. (1986). Towards a validated analysis of scientific text structure. *Applied Linguistics 7,* 1, 57-70.

Crookes, G. & Rulon, K. A. (1985). *Incorporation of corrective feedback in native speaker/non-native speaker conversation.* Technical Report no. 3, Center for Second Language Classroom Research, Social Science Research Institute, University of Hawaii.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika 16,* 407-424.

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,* (pp. 23-38). Washington, D. C.: TESOL.

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

Faraone, S. V. & Dorfman, D. D. (1988). Testing the significance of interobserver agreement measures in the presence of autocorrelation: the jackknife procedure. *Journal of Psychopathology and Behavioral Assessment 10,* 1, 39-47.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76,* 5, 378-382.

Fleiss, J. L., Cohen, J. & Everitt, B. S. (1969). Large sample standard errors of *kappa* and weighted *kappa*. *Psychological Bulletin* 72, 323-327.

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*, (pp. 115-161). Norwood, New Jersey: Ablex.

Frame, R. E. (1979). Interobserver agreement as a function of the number of behaviors recorded simultaneously. *The Psychological Record 29*, 287-296.

Freeman, D. (1983). *Margaret Mead and Samoa: the making and unmaking of an anthropological myth*. Cambridge, MA: Harvard University Press.

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

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

Gelfand, D. M. & Hartmann, D. P. (1975). *Child behavior analysis and therapy*. New York: Pergamon.

Guetzkow, H. (1950). Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology 6*, 47-58.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis 10*, 1, 103-116.

Hartmann, D. P. (1982). Assessing the dependability of observational data. In D. P. Hartmann (Ed.), *Using observers to study behavior* (pp. 51-67). San Francisco, CA: Jossey-Bass.

Hawkins, R. P. (1982). Developing a behavior code. In D. P. Hartmann (Ed.), *Using observers to study behavior* (pp. 21-36). San Francisco, CA: Jossey-Bass.

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

Hoyt, C. (1941). Test reliability estimated by the analysis of variance. *Psychometrika 6*, 153-160.

Judd, C. M. & Kenny, D. A. (1982). Research design and research validity. In D. Brinberg & L. H. Kidder, (Eds.), *Forms of validity in research*, (pp. 23-39). San Francisco: Jossey-Bass.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: the ABCs of reliability. *Journal of Applied Behavior Analysis 10*, 1, 141-150.

Kazdin, A. E. (1982). Observer effects: reactivity of direct observation. In D. P. Hartmann (Ed.), *Using* *observers to study behavior* (pp. 5-20). San Francisco, CA: Jossey-Bass.

Kirk, J. & Miller, M. L. (1986). *Reliability and validity in qualitative research.* Beverly Hills, CA: SAGE.

Kliess, H. O. & Bloomquist, D. W. (1985). *Psychological research methods*. Boston: Allyn & Bacon.

Kreckel, M. (1981). *Communicative acts and shared knowledge in natural discourse.* London: Academic Press.

Lahey, M. A., Downey, R. G. & Saal, F. E. (1983). Intraclass correlations: there's more than meets the eye. *Psychological Bulletin* 93, 3, 586-595.

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.

Mash, E. J. & McElwee, J. (1974). Situational effect on observer accuracy: behavioral predictability, prior experience, and complexity of coding categories. *Child Development* 45, 367-77.

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

Mead, M. (1928). *Coming of age in Samoa.* New York: Morrow.

Medley, D. M. & Mitzel, H. (1958). An application of the analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. *Journal of Experimental Education* 27, 23-35.

Medley, D. M. & Mitzel, H. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247-328). Chicago, IL: Rand-McNally.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin* 86, 376-390.

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

Nunnally, J. C. (1978). *Psychometric theory, second edition.* New York: McGraw-Hill.

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.

Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

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*, (pp. 99-126). New York: Plenum.

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

Ramirez, A. (1988). Analyzing speech acts. In J. L. Green & J. O. Harker (Eds.), *Multiple perspective analyses of classroom discourse*, (pp. 135-163). Norwood, New Jersey: Ablex.

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*, (pp. 147-165). Englewood Cliffs, New Jersey: Prentice-Hall International.

Renne, C. M., Dowrick, P. W. & Wasek, G. (1983). Consideration of the participant in video recording. In P. W. Dowrick & S. J. Biggs (Eds.), *Using video*, (pp. 23-32). New York: Wiley.

Rojahn, J. & Kanoy, R. C. (1985). Toward an empirically based parameter selection for time-sampling observation systems. *Journal of Psychopathology and Behavioral Assessment 7*, 2, 99-120.

Rosenthal, R. (1976). *Experimenter effects in behavioral research*. Enlarged edition. New York: Irvington.

Rosenthal, R. (1982). Valid interpretation of quantitative research results. In D. Brinberg & L. H. Kidder (Eds.), *Forms of validity in research*, (pp. 59-75). San Francisco: Jossey-Bass.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal 13*, 1, 51-59.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin 57*, 416-428.

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, (pp. 46-72).