The Development and Use of Classroom Observation Instruments

Jack Martin
simon fraser university

On a brièvement retracé l'historique et précisé l'utilisation des instruments de mesure permettant (1) l'observation du comportement des élèves et des professeurs en salle de classe et (2) la classification des diverses modalités de ce comportement pour une analyse ultérieure. On a défini ces instruments de mesure et énoncé ses principales caractéristiques. On y traite aussi en détail de l'approche séquentielle qui a présidé à l'élaboration de ces instruments. En conclusion, on élargit le propos et on fait la critique de ce programme de développement.

Classroom observation instruments are organized, objective systems for observing, coding, arranging, and analysing the behaviors emitted by teachers and students engaged in instructional exchanges. By breaking up the continuous stream of teacher–student behaviors into readily observable units, such instruments permit the production of permanent "shorthand" records of classroom interactions. These records can be analysed by teachers and research workers in a variety of ways and for a variety of purposes.

The focus of this paper is confined to *category* observation instruments. These instruments typically record classroom behaviors in the form of tallies, checks, or other marks which code them into predefined categories and yield information about which behaviors occurred and how often they occurred during the period of observation. *Rating* schemes, in which classroom observers rate the teacher, class, or pupils on one or more dimensions (even though these ratings may be based on direct observation of specific behaviors), are excluded from the discussion. *Sign* observation systems, in which an observer simply checks each behavior that occurs regardless of its frequency of occurrence, are similarly excluded. (Information pertaining to ratings and signs can be located in Medley & Mitzel, 1963; Ober, Bentley, & Miller, 1971; and Remmers, 1963.)

The major structural components of a category observation instrument are (1) a set of operationally defined categories of behavior; (2) a set of rules and priorities for observation and coding; (3) a standardized recording form (matrix, grid, etc.); and (4) a series of instructions for organizing and analysing the observational data. The first two components are absolutely necessary, whereas the final pair, while increasing the overall utility of an instrument, may at times be omitted.

The earliest category observation instruments developed specifically for the study of classroom behaviors appeared in the thirties and forties (Anderson & Brewer, 1945, 1946; Withall, 1949; Wrightstone, 1934). Since then, hundreds have no doubt been developed for various purposes pertaining to classroom instruction. Of these, only a handful have received systematic attention and have enjoyed widespread popularity and use. Flanders's (1960) system of classroom interaction analysis is undoubtedly the forerunner in this latter group. Other instruments of note have been developed by Amidon and Hunter (1966); Bellack, Kliebard, Human, and Smith (1966); Gallagher and Aschner (1963); Ober (1970); Hough and Ober (Note 1); and Smith and Meux (Note 2).

Category observation instruments have been used by educators in attempts to measure pupil participation, effective teaching behavior, classroom climate, and multiple dimensions of classroom interaction (cf. Medley & Mitzel, 1963). In recent years, these instruments have demonstrated their utility in teacher effectiveness research and in the tremendously important area of teacher education (cf. Flanders, 1970; Borg, Kallenbach, Kelley, & Langer, Note 3). The expanding panoply of functions has resulted in the employment of classroom category instruments in a wide variety of environments, not all of which are directly receptive to one or more of the carefully developed instruments which are currently available. Consequently, increasing numbers of educators have attempted to develop their own category systems to fit what they perceive as the unique features of their classrooms, research projects, or social-vocational milieux. Unfortunately, much of this work has proceeded in the absence of a thorough working knowledge of the criteria which category observation instruments must meet before the data they produce can be considered valid and reliable. Indeed, professional papers and programs based on data from faulty or carelessly employed category instruments are becoming alarmingly frequent. Good category observation systems are difficult, time-consuming, and often costly to develop. While it may be tempting to bypass this largely tedious process and still appear to reap the unique benefits of systematic methods of observation (Crano & Brewer, 1973), this is an extremely irresponsible tactic in which short-term "benefits" are likely to become longer-term "dead ends."

ESSENTIAL PROPERTIES OF THE CATEGORY OBSERVATION INSTRUMENT

A sound category observation instrument must be *objective, relevant, parsimonious, efficient, reliable,* and *valid.* While the maintenance of some of these criteria increases the likelihood that others will be realized, all will be viewed independently for discussion purposes.

Objectivity (specificity) simply means that the various categories which the instrument comprises be defined so that the definitions are directly reducible to empirical realities. Stated another way, behavioral categories

must possess the status of intervening variables, not hypothetical constructs. This operationism ensures that the behaviors recorded are directly observable and prepares the way for reliable coding across a number of independent observer-recorders.

The criterion of relevance helps determine the correspondence of the observation instrument to the classroom behaviors it purposes to pinpoint. If the purpose of a category system is to study behaviors, A, B, and C, these behaviors must have a reasonable likelihood of occurring in the classroom contexts being observed. One would presumably not employ an instrument developed for studying disruptive student behaviors at the elementary level to analyse the interactions taking place in post-graduate seminars. The property of relevance is thus closely linked to the purpose for which the instrument is devised. Just as objectivity lays the groundwork for interrater reliability, relevance prepares the way for instrument validity. (It should be noted that the criteria being discussed here should determine selection of existing instruments, as well as development of new ones.)

Parsimony is a third criterion, for complexities in instrument construction can severely limit the practical utility of the category system. The number of categories which an instrument comprises should be relatively small. The construction of recording grids, and indeed the coding procedures themselves, should ensure a reasonable degree of simplicity in the recording process. Fancy, theoretically based frills should always be avoided. The task of recording is difficult enough without asking the observer-recorder to make esoteric deductive decisions and to transfer these to a labyrinth of intersecting squares. The observer-recorder, to do the job adequately, must be assisted by the instrument engineer. The role of observer should be that of descriptive stenographer, not inferential pedant. Adherence to the criterion of simplicity can save a tremendous amount of time and money when it comes to training observers to use the category instrument.

Efficient category systems are those in which all categories are employed in the recording process and practical discriminations between categories are relatively salient. Decisions as to which behavior goes where should be easy for the observer-recorder to make after a reasonable amount of training in the use of the instrument. When the system is applied to the general classroom setting for which it was developed, the pattern of recording tallies should be diffused across all category areas. The criterion of efficiency, by guarding against the capricious construction of ambiguous and seldom-employed categories, ensures that existing categories serve a legitimate recording function.

The final criteria, reliability and validity, are essential to any category observation instrument just as they are to quantification devices of any kind. Medley and Mitzel (1963) in their classic article "Measuring Classroom Behavior by Systematic Observation" define the reliability of a category observation instrument as "the extent that the average difference

between two measurements independently obtained [e.g., by two separate observer-recorders] in the same classroom is smaller than the average difference between two measurements obtained in different classrooms" (p. 250). Unreliability, indicated when two measures of the same class differ too much, can occur because "the behaviors are unstable, because the observers are unable to agree upon what occurs, because the different items which enter into the measurement lack consistency, or for some other reason" (Medley & Mitzel, 1963, p. 250).

Validity is defined by the same authors as "the extent that differences in scores yielded by it [the category observation instrument] reflect actual differences in behavior – not differences in impressions made on different observers" (p. 250). There are, of course, several varieties of validity, but generally speaking, reliability is always necessary before an instrument can be valid. Reliability is, however, not sufficient to ensure validity.

A detailed discussion of reliability and validity is not attempted here. The important point to be emphasized is that any category observation instrument will very likely meet these two final criteria if the instrument developer takes sufficient pains to ensure that his instrument is objective, relevant, parsimonious, and efficient. A sequential program for developing observation instruments which possess these properties is the focus of this paper.

DEVELOPING THE CATEGORY OBSERVATION INSTRUMENT

1. Descriptive Observation

The first problem in constructing a category observation instrument is deciding which behaviors are to be recorded. Obviously, such decisions cannot be separated from the purposes for which a particular instrument is being developed. Such purposes normally arise from a variety of theoretical considerations. While *a priori* judgments as to which behaviors should be noted are thus impossible to avoid completely, the criteria of objectivity and relevance demand that they be kept to an absolute minimum. One of the best means of accomplishing this end is to begin by conducting informal *descriptive observations* of classroom interactions typical of those to which the instrument will eventually be applied. Ideally, observers unaware of the instrument developer's theoretical biases should independently monitor such interactions by simply writing down a series of descriptive phrases (in everyday language) which depict the behavioral phenomena they witness. A pool of descriptive phrases from a representative sampling of classroom interactions provides a reasonably sound inductive basis upon which to build a category instrument. By confining himself/herself to this pool, the instrument developer can ensure that the categories will be well rooted in the empirical realities of the classroom (objectivity), and that the specific behaviors which will eventually define the categories are commonly displayed (relevance). Audio- and videotape recordings of classroom interactions can be very useful during the initial

phase of descriptive observation (and indeed throughout all the developmental steps). The permanent records of classroom behaviors they provide can be viewed many times, in many ways, and by many different observers.

2. Initial Unitizing and Categorizing

Perhaps the most crucial decisions that category instrument developers must make are those of *unitizing* and *categorizing*. Such decisions are almost impossible in the absence of the descriptive observation data collected in the first phase of development. Categorizing refers to the process of grouping informal observations into a number of generic blocks which are meaningful in relation to the purposes of the instrument developer. Such blocking normally proceeds in stages, the first and most obvious of which is the grouping together of relatively synonymous behavioral descriptions. This initial grouping, while congruent with the semantics of a particular verbal community, is quite free of specific *a priori* theorizing. Once synonymous blocking has been completed, further grouping or categorizing reflects more directly the dimensions of interaction deemed to be important by the instrument developer. These dimensions are largely determined by individual theoretical and/or empirical proclivities. While a certain amount of deduction is obviously involved, the experienced instrument developer makes considerable use of the empirical findings of educational research and classroom interaction analysis which bear directly upon the phenomena he/she purposes to investigate. The more such deductions coincide with the inductive work of others in the chosen area of investigation, the less likely a category constructor will waste time "reinventing the wheel."

The process of categorizing is basically one of successive approximation in which numerous behavioral descriptions are gradually channelled into a finite number of generic behavioral blocks. These categories, when taken together, should be a parsimonious abridgment of the variance and substance of the initial "raw" observations. The final number of categories should not be too large; few studies have used more than a dozen. Nonetheless, it is probably preferable, at this stage, to have too many rather than too few categories. Further empirical testing can always throw two categories together if this is desirable. It is much more difficult to create new categories later in the developmental sequence.

To ensure the efficiency of the instrument, categories should be defined so that the number of descriptive observations (i.e., those collected in step 1) fitting into each is roughly the same. (While this is a sound general rule, infrequently employed categories are sometimes desirable if the discriminations they permit are crucial to the purposes of the instrument user.) If a category system contains a category such as "Miscellaneous," "Other," or "Unclassified," the number of behaviors falling into this category should be small. At later stages in the development of the instrument, coding frequencies in this category should be correspondingly low.

Unitizing refers to the process of defining the standard behavioral unit which is to be recorded. The instrument developer must determine the most reasonable means of breaking up the continuous stream of classroom behaviors into recognizable units which can be readily and reliably noted by trained observers. The unit of behavior to be tallied may be a natural one (e.g., a question, a simple sentence or its nonverbal equivalent, a contact) or a brief time unit. When a natural unit is used, each tally represents an occurrence; when a time unit is used, each tally represents a period of time which includes at least one occurrence (Medley & Mitzel, 1963). Whatever the nature of the behavioral unit, it should be small (or short) enough so that observers do not code in terms of retrospective impressions or general inferences. Observers should be kept continuously busy (at a reasonable and consistent rate) so that their personal inference-making does not interfere with the objective-descriptive process. If observers have time to code on the basis of their general impressions, the data obtained from the category observation instrument will be distorted in terms of reliability, validity, or both (cf. Medley & Mitzel, 1963, p. 305 – the "halo-effect").

It is important to remember that the behavioral unit selected must be consistently employed across all categories. In other words, it is not permissible to code some categories in terms of natural events and others in terms of time. Further, if natural units are employed, they should be relatively uniform across all category areas. The advantage of natural units is that they produce results which are easier to interpret. The number of smiles a teacher emits in an hour class period is probably more meaningful than the number of 3-second periods in which her facial expression was predominantly that of "smiling." Time units, however, have the advantage that they can be more consistently employed across a number of observers and over a number of categories.

3. Pilot Testing and Cleaning

After the instrument developer has made initial decisions of categorizing and unitizing, he/she is normally required to subject this observation instrument to a series of pilot tests. The object of such tests is to gradually optimize the reliability and validity of the instrument. If reliability and validity coefficients are unsatisfactory, the category observation instrument must be re-examined in terms of the basic criteria of objectivity, relevance, parsimony, and efficiency. The instrument can then be altered (*cleaned*) in directions suggested by these criteria and the revised instrument subjected to an additional set of pilot tests. This cyclic process of field-testing and cleaning continues until satisfactory reliabilities and validities have been demonstrated across a number of trained observers and a sampling of classroom situations representative of those for which the instrument is being designed.

A very important part of the "pilot testing and cleaning" step is the

training of observer-coders. The amount and difficulty of training is a direct function of the ease of discrimination between the various categories within the instrument. The lower the coefficient of observer agreement (i.e., the correlation between scores based on observations made by different observers at the same time), the more difficult the inter-category discriminations, and consequently the less objective the instrument. Perhaps the most comprehensive method of calculating precise coefficients of reliability is the analysis of variance technique reported by Medley and Mitzel (1963). However, if this procedure goes beyond the statistical knowledge of the instrument developer, or exceeds available time parameters, coefficients of observer agreement can be reasonably, although more generally, calculated with a variety of statistics such as chi-square or Scott's coefficient of reliability. These latter statistics are commonly employed in observer training and research programs.

Objectivity, which is dependent upon the difficulty of the judgments required and the degree of observer sophistication and training, can be increased by extending and clarifying the behavioral bases for category discriminations. A good deal of time should be spent specifying, in simple empirical terms, the behavioral cues to which the observers must respond. If these cues cannot be found, category discriminations should be "eclipsed" or moved to other behavioral dimensions. An example of eclipsing would be the combining of two or more categories with a seemingly homogeneous behavioral base into a single category. The alternative strategy consists of emphasizing behavioral distinctions not presently highlighted, thus forcing a discrimination where none was previously possible. This second approach often involves relabelling or renaming existing categories. Which approach is most desirable must, of course, be determined by the objectives of the instrument developer. In making such decisions, the instrument developer should remember that required discriminations must be relevant to the classroom interactions being studied and must preserve the parsimony and efficiency of the instrument as a whole. The process of category cleaning ensures that discriminations are based on actual differences in behavior. General dimensions such as "genuineness" and "honesty" may seem distinguishable on an *a priori* basis, but discernible behavioral differences between the two may be forced or overly impressionistic.

The sequence of testing and cleaning should be repeated until the behavioral cues upon which category discriminations rest are explicit even to the uninitiated, and easy for the trained observers. Repetitive calculations of coefficients of observer agreement can serve to document progress towards these objectives.

4. Formal Categorizing and Prioritizing

Step 4 is essentially concerned with formalizing the revisions made to the category observation instrument during the preceding developmental phase. Category rubrics should be scrutinized to ensure that they directly

reflect the behavioral phenomena which they encompass; and any inter-category discriminations which are not immediately obvious to trained observers should be explicitly formalized in a set of coding priorities. The first of these tasks is generally quite easy if previous developmental steps have been carefully conducted. The second is more difficult and consequently deserves further attention.

Prioritizing refers to the process of developing rules on the basis of which an observer-coder can record "more important" behaviors at the expense of "less important" behaviors during complex or rapid interaction segments which make exhaustive recording impossible. Prioritizing is always necessary when category observation instruments are employed, since such instruments ideally attempt to record all classroom behaviors under relevant category headings. Any observer of interpersonal interactions can attest to the fact that interaction rates are not constant across time. The observer-coder is periodically inundated with peak periods of activity which are extremely difficult to record, especially since these often follow upon periods of relative lethargy during which the recording process is quite simple. To assist the observer in recording peak activity periods, it is necessary to formulate rules of recording which specify those behaviors which must be recorded and those which can be ignored given insufficient time to note everything that occurs. Typically, prioritizing involves a listing of all categories in terms of their relative importance vis-à-vis the objectives of the instrument developer. By familiarizing themselves with this priority list, two or more observer-coders can reliably record periods of peak activity even though neither records all the behaviors which occur.

In addition to the main priority list, priority rules should be developed for those few interaction sequences which are ambiguous given existing category boundaries. Even when a category instrument has been thoroughly cleaned, situations will undoubtedly arise in which behavioral discriminations are not completely obvious. Interaction sampling during pilot tests can never exhaust all interaction possibilities. A few troublesome interactions will need to be coded in terms of standardized rules. Thus, prioritizing usually includes the formulation of a few specific coding rules as well as a general rule of category priority. While priority rules are an essential component of any category observation instrument, they should normally be few in number. Prioritizing should not be used to compensate for inadequate pilot testing and category cleaning.

The process of prioritizing is obviously dependent upon whether behavioral recording is conducted (1) on the basis of time units or event units, (2) with or without audio- or videotape assistance, and (3) by one or more than one observer. If the recording task is to be accomplished *in vivo*, priority rules are much more necessary than if the observer-coder records from videotape, which can be stopped, reversed, and replayed. Similarly, if two or more observers divide the observation task, each taking respon-

sibility for a subset of the total number of categories, priority rules are less essential. Prioritizing, like categorizing and unitizing, is dependent upon the recording procedures a category instrument developer chooses to employ. While priority rules often suggest themselves quite naturally from the previous phase of pilot testing and cleaning, many experienced instrument developers examine the utility of their priority rules by conducting a second series of pilot tests at this stage.

5. Standardizing

The final phase in the present developmental sequence consists of standardizing the use of the category observation instrument. The procedures associated with a category instrument for recording classroom behaviors, analysing behavioral records, and training observer-coders to perform these tasks should be clearly stated.

The various categories in the observation instrument should be arranged on a recording schedule or grid (normally a matrix on a sheet of paper of appropriate size) in a manner which facilitates ease in recording. It is preferable, in most instances, to isolate different behavioral categories along the various rows of this matrix so that different behaviors can be identified by the position of the tally marks in the matrix rather than their form; it is easier to record different behaviors in different spaces using the same symbol than it is to switch among different symbols. With a bit of training, the observer is able to locate the proper spaces without consciously attending to the recording matrix itself. If it is important to obtain information about the pattern of classroom behaviors across time, in addition to their actual frequencies, columns can be constructed which intersect the various category rows. The observer moves one column to the right after each tally and thus records the sequence of the interaction pattern. Whatever the recording system eventually devised, it should be kept constant, and a standardized scoring grid should accompany the instrument.

The instrument developer should also indicate how the observational data can be grouped together and analysed in meaningful ways. Calculations of raw and relative frequencies, emission percentages, and totals and subtotals by time and category are standard quantitative manipulations. In addition, some category system developers have documented the calculation of standard interaction indices and ratios (cf. Bales, 1950, 1970; Flanders, 1963) which permit meaningful summations of a great many behavioral codings in terms of underlying theoretical and empirical positions. The use of standard indices encourages easy comparison of results from different investigations and situations. Instrument developers should attempt, in standardizing their scoring and analysis procedures, to foster this kind of informational interchange.

Finally, the development of standard procedures for training observer-coders in the use of the category observation instrument is an important consideration. Training procedures are extremely difficult to develop,

since our present knowledge in this area is rather miniscule. Consequently, many instrument developers have ignored systematic development in this area. While it is beyond the scope of this paper to detail specific training strategies, the question of observer training is an exceedingly crucial one. Those training programs which seem to have been most successful (from the author's own experience) are typically based, to a greater or lesser extent, upon what might be called a task analysis of the activity of sophisticated observer-coders. Simply stated, this procedure involves determining the subsets of behaviors which make up the total observation-recording process, isolating these subsets, and sequencing their acquisition from simple to complex. The successive approximation strategy implicit in such training programs is an extremely effective learning paradigm (cf. Becker, Englemann, & Thomas, 1971).

CONCLUDING REMARKS

The formation of standard procedures for recording, analysing, and training concludes the developmental sequence. While there are probably other methods of instrument construction which are equally useful, the one just delineated, if followed closely, ensures that the various criteria upon which a category observation instrument might be adjudicated (objectivity, relevance, parsimony, efficiency, reliability, and validity) will be satisfactorily met. Some of the steps will obviously overlap and interact much more dynamically in practice than they do on paper. Consequently, the developmental sequence discussed should be considered a branching, rather than a strictly linear, program. Pilot testing, for example, no doubt has a place in each phase and should not necessarily be limited to the third.

A second caveat to the foregoing text concerns its failure to discuss several important dimensions which must be crucial elements in the thinking of category instrument developers. The dimension running from induction to deduction was considered briefly, and the importance of induction stressed as it relates to objectivity and relevance. A second dimension, not discussed, is that of structuralism–functionalism. In the past, most category observation instruments have recorded classroom behaviors on the basis of their formal structural properties. Recently, some instrument developers (Martin, 1975) have attempted to record behaviors in terms of their effects or functions (i.e., what occurs after they are emitted) rather than their formal response properties, arguing that the meaning of any behavior resides not in its static structure but in its dynamic relationship to preceding and to subsequent events. If a teacher asks a question, the question is important only insofar as it occasions a particular kind of pupil response and should be recorded on the basis of its effect upon student(s). What is structurally a high-order question does not function as a high-order question if the pupil to whom it is addressed does not respond appropriately. Functionalists argue that by recording classroom behaviors structurally it is possible to obtain a very warped and misleading

picture of what is actually occurring in the classroom and how this activity relates to the basic processes of teaching and learning.

Still another dimension of importance to the category instrument developer is that of process versus content. While this dimension overlaps with the previous one, it highlights the importance of "frame of reference" as in "who speaks to whom." Most category systems do not attempt to code the direction of communications, but obviously this can be an important variable. If a teacher reacts to Jimmy's statement in an attempt to have him extend it, but Sally interrupts before Jimmy can respond, the actual process of interaction (and perhaps learning) is different from the intended interaction in which Jimmy would have responded and Sally would have remained silent. This would be true even if the contents of the communications in these two instances were identical. Most category observation instruments are not constructed to accurately pick up such process differences. (Bales's system of Interaction Process Analysis records who-to-whom patterns, but it is not, strictly speaking, a *classroom* observation instrument.)

Students of classroom interaction are becoming increasingly sophisticated in their thinking about, and in their efforts to examine, classroom behavior. Obviously, no one category instrument can be all things to all users. It is not a shortcoming for an instrument developer to ignore various poles in dimensions such as those just discussed, if the developer clearly states the objectives of his/her instrument and its areas of application.

Classroom observation instruments have proven to be extremely useful methodological tools in many areas of education, from research on teacher effectiveness to pre- and in-service teacher training programs. Observation methods permit direct examinations of the behaviors of teachers and students in regular classroom situations. Perhaps more can be learned about critical variables in teaching and learning from the data obtained by observation instruments than from any other single source. Given the tremendous promise of such instruments, it is exceedingly important that they be developed and employed in ways which will ensure that the information they yield will be both reliable and valid.

NOTE

An earlier draft of this manuscript appeared in volume 12, no. 1, of the *Journal of Classroom Interaction*.

REFERENCE NOTES

1. Hough, J., & Ober, R. The effect of training in interaction analysis on the verbal behavior of pre-service teachers. Paper read at American Educational Research Association annual meeting, Chicago, February, 1966.

2. Smith, B. O., & Meux, M. O. *A study in the logic of teaching*. U.S. Office of Education, Cooperative Research Project No. 258 (7257), Urbana, 1962.

3. Borg, W. R., Kallenbach, W. W., Kelley, M. L., & Langer, P. The minicourse: A rationale and uses in the in-service education of teachers. Paper read at American Educational Research Association annual meeting, Chicago, February, 1968.

REFERENCES

Amidon, E. J., & Hunter, E. *Improving teaching: Analyzing verbal interaction in the classroom.* New York: Holt, Rinehart and Winston, 1966.

Anderson, H. H., & Brewer, H. M. Studies of teachers' classroom personalities. I. Dominant and socially integrative behavior of kindergarten teachers. *Applied Psychological Monographs*, 1945, no. 6.

Anderson, H. H., & Brewer, H. M. Studies of teachers' classroom personalities. II. Effects of teachers' dominative and integrative contacts on children's classroom behavior. *Applied Psychological Monographs*, 1946, no. 8.

Bales, R. F. *Interaction process analysis.* Reading, Mass.: Addison-Wesley, 1950.

Bales, R. F. *Personality and interpersonal behavior.* New York: Holt, Rinehart and Winston, 1970.

Becker, W. C.; Englemann, S.; & Thomas, D. R. *Teaching: A course in applied psychology.* Chicago: Science Research Associates Inc., 1971.

Bellack, A. A.; Kliebard, H. M.; Human, R. T.; & Smith, F. L., Jr. *The language of the classroom.* New York: Teachers College Press, 1966.

Crano, W. D., & Brewer, M. B. *Principles of research in social psychology.* New York: McGraw-Hill, 1973.

Flanders, N. A. *Teacher influence, pupil attitudes and achievement.* Minneapolis: University of Minnesota, 1960.

Flanders, N. A. Intent, action and feedback: A preparation for teaching. *Journal of Teacher Education*, 1963, *14*, 251–260.

Flanders, N. A. *Analyzing teaching behavior.* Reading, Mass.: Addison-Wesley, 1970.

Gallagher, J. J., & Aschner, M. J. A preliminary report on analysis of classroom interaction. *Merrill-Palmer Quarterly of Behavior and Development*, 1963, *9*, 183–194.

Martin, J. *Classroom process analysis: A training manual for student teachers.* Burnaby, B.C.: Simon Fraser University, 1975.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: American Educational Research Association, Rand McNally Inc., 1963.

Ober, R. L. The reciprocal category system. *Journal of Research and Development in Education*, 1970, *4*, 34–51.

Ober, R. L., Bentley, E. L., & Miller, E. *Systematic observation of teaching: An interaction analysis–instructional strategy approach.* Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Remmers, H. H. Rating methods in research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: American Educational Research Association, Rand McNally Inc., 1963.

Withall, J. Development of a technique for the measurement of socio-emotional climate in classroom. *Journal of Experimental Education*, 1949, *17*, 347–361.

Wrightstone, J. W. Measuring teacher conduct of class discussion. *Elementary School Journal*, 1934, *34*, 454–460.

Jack Martin is an assistant professor in the Faculty of Education at Simon Fraser University, Burnaby, B.C., V5A 1S6.