DOCUMENT RESUME

ED 137 371 95 TM 006 172

AUTHOR Borich, Gary; And Others

TITLE Classroom Observation Data: Is It Valid? Is It Generalizable? A Compendium of Methodological Papers.

INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.

SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.

PUB DATE [Apr 77]

CONTRACT NIE-C-74-0088

NOTE 92p.; Papers presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.

DESCRIPTORS *Academic Achievement; *Classroom Observation Techniques; Classroom Research; Effective Teaching; Elementary Education; Junior High Schools; Mathematical Models; Matrices; Reading Achievement; Reading Instruction; *Reliability; Student Behavior; *Student Teacher Relationship; *Teacher Behavior; *Validity; Video Tape Recordings

IDENTIFIERS CERLI Verbal Behavior Classification System; Classroom Communication Observation System; Flanders Interaction Analysis; Generalizability Theory; Observation Schedule and Record; Spaulding Teacher Activity Rating Schedule; Teacher Child Dyadic Interaction System (Brophy)

ABSTRACT

The issues discussed in these four papers concern the validity and generalizability of classroom observation instruments. These issues have been studied and are reported here in an attempt to better define the limits to which classroom observation instruments can be used in researching relationships between teacher behavior and student outcome. The premise undergirding these investigations is that before consistent and positive process-product relationships can be found, investigators must be cognizant of the sources of variance which affect the validity and generalizability of their process measures and which, in turn, affect the credibility of their research findings. The four papers are: "Convergent and Discriminant Validity of Five Classroom Observation Systems: Testing the Model" by G. Borich, D. Malitz, C.L. Kugle, and M. Pascone; "Generalizability of Teacher Behaviors Across Classroom Observation Systems" by D. Calkins, G. Borich, M. Pascone, and C.L. Kugle; "Measuring Classroom Interactions: How Many Occasions Are Required to Measure Them Reliably?" and "Generalizability of Teacher Process Behaviors During Reading Instruction" both by O. Erlich and G. Borich. (RC)

CLASSROOM OBSERVATION DATA:

Is it valid? Is it generalizable?

A Compendium of Methodological Papers

Gary Borich
Dick Calkins
David Malitz
Oded Erlich
Cherry Kugle
Maria Pascone

Evaluation of Teaching Project
The Research and Development Center for Teacher Education
The University of Texas at Austin
Austin, Texas

CLASSROOM OBSERVATION DATA:

Is it valid? Is it generalizable?

A Compendium of Methodological Papers

Gary Borich

Dick Calkins

David Malitz

Oded Erlich

C. L. Kugle

Maria Pascone

Evaluation of Teaching Project

The Research and Development Center for Teacher Education

The University of Texas at Austin

Austin, Texas

Papers presented at the American Educational Research Association annual meeting, New York City, April 4-8, 1977.

These studies were supported in part by National Institute of Education Contract NIE-C-74-0088, Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred.

Preface

The following papers represent modest attempts to bring clarity to a complex problem. The issues discussed in the following four papers concern the validity and generalizability of classroom observation instruments. These issues have been studied and are reported here in an attempt to better define the limits to which classroom observation instruments can be used in researching relationships between teacher behavior and student outcome. The premise undergirding these investigations is that before consistent and positive process-product relationships can be found, investigators must be cognizant of the sources of variance which affect the validity and generalizability of their process measures and which, in turn, affect the credibility of their research findings.

GDB

TABLE OF CONTENTS

Convergent and Discriminant Validity of Five Classroom Observation Systems: Testing the Model

Gary D. Borich, David Malitz,
C. L. Kugle, & Maria Pascone

The University of Texas at Austin

Numerous instruments have been developed to systematically observe classroom behavior. These instruments typically consist of a number of categories of teacher-student behavior which an observer tallies or rates periodically as he watches classroom interaction. For the greater part of a decade researchers have used such instruments to investigate the relationship between teacher behavior and student outcome, but this effort has yielded relatively few consistent findings.[1,2] While many possible reasons for the dirth of consistent findings can be advanced, two which must be considered are that the research model or theory implicit in process-product investigations may be inadequate or too simplistic to uncover such relationships and psychometric weaknesses within instruments used by the researchers may obscure any underlying relationships which do exist.

At present, there is no a priori reason to suspect one of these possibilities over any other. However, as process-product studies themselves confirm (Brophy & Evertson, 1976; Good & Grouws, 1975, McDonald et al., 1975; Stallings & Kaskowitz, 1974) there has been a conspicuous lack of validity studies of the research instruments used, especially instruments to measure teacher behavior. Taking note of this Borich (1977) detailed several of the most salient sources of invalidity afflicting observational measures of teacher behavior, but did not provide empirical data as to the actual effect of these sources of invalidity on instruments used to measure teacher behavior. The present study undertook to determine the extent to which one of these sources of invalidity, the lack of convergent and discriminant validity, was present in five classroom observation systems. The validity model reported by Campbell and Fiske (1959) was employed: this model requires that both convergent and discriminant validity be demonstrated.

Convergent validity is a confirmation of traits (or variables or categories) by independent measuring methods that requires significant correlation between two methods (or systems) measuring the same trait. Discriminant validity is a requirement that "the correlation between different measures measuring the same trait exceed (a) the correlations obtained between that trait and any other trait not having method in common and (b) the correlations between different traits which happen to employ the same method" (Borich & Malitz, 1975). By determining intercorrelations among categories in a multitrait-multimethod matrix, one can identify categories which pass specified tests of convergent and discriminant validity. These procedures were applied to the following data in order to ascertain the external validity of five classroom observation systems.

Method

The data were obtained from videotapes of twelve in-service junior high school teachers, each teaching the same content, a unit in social studies, on three occasions of approximately 50 minutes duration. Each of 36 videotapes was rated by five pairs of coders, each pair trained in a different observational coding system. For two of the five observational systems, coders were employed who had previously been trained by the authors of these systems, these being the two most complex systems. The remaining three pairs of coders were trained by the investigators from training materials supplied by system authors and from standard protocols from system manuals.

The five systems employed for this study were selected from Simon and Boyer's Mirrors for Behaviors (1970). The systems were (1) the Observation Schedule and Record, OScAR 5, (Medley & Mitzel, 1959), (2) Spaulding Teacher Activity Rating Schedule, STARS (Spaulding, 1967), (3) Flanders System of Interaction Analysis, (Flanders, 1971), (4) CERLI Verbal-Behavior Classification System, CVC (Cooperative Educational Research Laboratory) (Note 1), and (5) The Classroom Communication Observational System, CCO (Withal, Lewis & Newell, 1961). These systems were selected because of their availability to the educational research community (and therefore presumed use) and for the number of categories and associated operational definitions they had in common--the latter being a requirement for assessing convergent validity.

Upon completion of training, system coders, using their respective systems, rated three trial videotapes of the same general form and content as the experimental tapes in order to obtain estimates of interjudge reliability prior to the study. While reliabilities varied due to system complexity, all were deemed acceptable and are reported in Table 1 along with the median reliability for each pair of coders over all 36 tapes.

Insert Table 1 about here

Descriptions of the behavior categories of the three systems were obtained from the coding manuals, and categories were grouped across systems, if from the category descriptions it appeared that they measured the same behavior. From these comparisons, two categories were paired across the Flanders and CCO systems, three categories were paired across the Flanders and OScAR systems, three categories were paired across the STARS and CVC systems, four categories were paired across the OScAR and CCO systems, two categories were paired across the STARS and OScAR systems, and six categories were paired across the CVC and OScAR systems, for a total of six two-system comparisons. In addition, there was one three-system comparison, two categories were compaired across Flanders, CCO, and OScAR. A description of the behaviors comprising these comparisons appears in Appendix A.

In certain cases, a single variable from one system was paried with several variables in another system. This procedure was most commonly employed when a subset of categories on one system was encompassed by a single general category on another and when members of the subset were coded independently of each other, i.e., were discreet behavioral categories. For example, in the Flanders vs. OScAR comparison, the total Flanders frequency in category 9, Student Talk-Initiation was correlated with the sum of the OScAR frequencies in categories 10, Pupil Nonsubstantive Utterance; 20, Pupil Question; 30, Pupil Statement; and 40, Pupil Response. Matches were not made, however, for which the meaning of a category on one system would have to be split among different categories on the other system, i.e., a category could not be applicable to more than a single category or homogeneous subset of categories on another system.

Once the categories to be investigated had been identified, Pearson product-moment correlations were computed. These correlations were used to construct seven multitrait-multimethod matrices. For each matrix, a heterotrait-heteromethod block was formed with those values in which categories may or may not coincide but systems differ. A heterotrait-heteromethod block is illustrated in Figure 1.

Insert Figure 1 about here

For each matrix, a diagonal (called the validity diagonal) is formed through the heterotrait-heteromethod block by the series of cells in which categories coincide but systems differ. Values in the validity diagonal which are significantly different from zero are evidence for convergent validity. Discriminant validity must be assessed in two steps. First, each validity value must be compared with all values in ıts row and column in the heterotrait-heteromethod block to determine whether the correlation between different methods of measuring the same category exceeds correlations between that category and other categories not having method in common. Second, the heterotrait-monomethod triangles are examined to determine whether the correlation between different methods of measuring the same category exceeds correlations between that category and other categories which have method in common. This step is completed by comparing each category's validity diagonal value with values in the heterotrait-monomethod triangles in which that category is involved. This two-step procedure was carried out for each validity diagonal value in each of the seven matrices, and the results entered in Tables 2-8. In Figure 1, the validity diagonal for category "A" is significant at the .05 level and, therefore, it can be taken as evidence for convergent validity. Also, category

"A" presents good evidence for discriminant validity, since its validity diagonal value exceeds all of the values specified in the two-step procedure outlined above. Category "B", on the other hand, indicates neither convergent nor discriminant validity.

Results

Seven matrices resulted from the process of comparing categories and groups of categories across the five systems. Six of these matrices compared categories across two systems. The seventh matrix involved three of the systems. No matching categories were found to exist across any four or all five of the systems. A category or group of categories which was found to match across two or more systems will be referred to as a comparison category (CC). Twenty-three such CC's were created and will be referred to by number (i.e., CC1 through CC23). Appendix A lists each of these 23 CC's and the constituent system categories which comprise each CC. Of the six, two-system matrices, five contained four or less CC's. The multitrait-multimethod (MTMM) matrices for these five, two-system comparisons are shown in Tables 2 through 6. The three-system matrix is presented in Table 7. One two-system matrix contained six CC's and is shown in Table 8. Since this matrix is somewhat cumbersome to evaluate in its raw form, a summary table, Table 9, was constructed to aid in its evaluation.

Table 2 shows the matrix resulting from the matching of two categories across the Flanders and CCO systems. It can be noted that both CC1 and CC2 pass the criterion for convergent validity since both CC's have significant validity diagonal values (.7699 and .6620 respectively, $r_{.05} = .325$, df = 35). Since both CC's pass the test for convergent validity, they may be examined for discriminant validity. It will be recalled that determining discriminant validity is a two-step process. The first step involves comparisons of each CC's validity diagonal value with the other values in its row and column in the heterotrait-heteromethod block. The second step requires comparison of the validity diagonal value for each CC with values in the heterotrait-monomethod triangles. For both CC1 and CC2, the validity value exceeds the heterotrait-heteromethod values. Thus, both CC's meet the first cirterion for discriminant validity, since in both cases the correlation between different methods of measuring the same behavior exceeds correlations between that category and other categories not having method in common. In addition, the validity value exceeds the heterotrait-monomethod values. In other words, the correlation between different methods of measuring the same behavior exceeds correlations between that category and other categories having method in common--the second step. In summary, CC1 and CC2 pass all tests for convergent and discriminant validity.

Insert Tables 2 & 3 about here

Table 3, contains the results of matching categories across the Flanders and OScAR systems. Three CC's resulted from this comparison, CC3, CC4, and CC5. Examination of the validity diagonal reveals evidence for convergent validity for CC3 and CC4 since their values are significant. CC5 has a nonsignificant validity value, and therefore need not be examined for discriminant validity. CC3 and CC4 pass both the first and second steps for discriminant validity, since their validity diagonal values exceed all relevant values in both the heterotrait-heteromethod block and in the heterotrait-monomethod triangles. Thus, CC3 and CC4 pass all tests for convergent and discriminant validity. CC5 lacks evidence for convergent validity and therefore its discriminant validity need not be examined.

Table 4 shows the three CC's (CC6, CC7, and CC8) resulting from comparison of the CVC and STARS systems. None of these three CC's have significant validity diagonal values and therefore lack evidence of convergent validity. Discriminant

Insert Table 4 about here

validity is also necessarily lacking and therefore need not be examined.

Table 5 indicates that four CC's (CC15, CC16, CC17, and CC19) resulted from comparison of the CCO and OScAR systems.

Insert Table 5 about here

Examination of the validity diagonal values indicates that only CC15 and CC19 have significant values. However, both CC15 and CC19 fail the first step in the assessment of discriminant validity since both are exceeded by the heterotrait-heteromethod value of .5054. While the validity values for these categories pass the second test by exceeding all values in the heterotrait-monomethod triangle, they do not pass the first test for discriminant validity. Thus, while CC15 and CC19 show evidence for convergent validity, they show mixed results for discriminant validity.

Table 6 shows the results of the comparison of STARS with OScAR. Two CC's (CC20 and CC21) resulted from this comparison and both pass all tests for convergent and discriminant validity.

Insert Table 6 about here

When the Flanders, CCO, and OScAR Systems were compared, two CC's were found. The results of the comparison of CC22 and CC23 are presented in Table 7. Analysis of a three-system matrix proceeds in exactly the same manner

Insert Table 7 about here

as the analyses of a two-system matrix, except that instead of one validity diagonal to examine, there are now three (corresponding to the three system pairings). Examination of the three validity diagonals indicates all values are significant. Furthermore, it can be noted that each of these values exceeds the relevant values in the heterotrait-heteromethod blocks and in the heterotrait-monomethod blocks. Thus, both CC22 and CC23 pass all tests for convergent and discriminant validity in the Flanders, CCO, and OScAR comparison.

Lastly, comparison of the CVC and OScAR systems resulted in the creation of six CC's (CC9 through CC14). The correlation matrix for these comparisons is shown in Table 8 while a summary table of these data are presented in Table 9. Table 9 shows the validity diagonal value for each CC. In addition, data are presented pertaining to each CC's discriminant validity (the highest value in the relevant parts of the heteromethod and monomethod blocks and the number of times

Insert Table 8 & 9 about here

the validity value is exceeded in each of these blocks). Examination of this table reveals that three CC's (CC10, CC12, and CC14) have non-significant validity values and therefore lack evidence for convergent validity. Of the remaining three CC's, all show good evidence of discriminant validity since their validity values are exceeded by none of the relevant heteromethod or monomethod values.

Comparison of the five teacher observation systems employed in this study produced 23 CC's. Twenty-one of these CC's were involved in two-system comparisons and two in a three-system comparison. Of the 23 CC's, 13 (57%) showed evidence of convergent validity. Eleven of the 23 CC's (48%) passed tests for both convergent and discriminant validity. Thus, of the CC's which were analyzed, only about half conformed to Campbell and Fiske's criteria for convergent and discriminant validity.

Discussion

The purpose of this research has been to evaluate the convergent and discriminant validity of five classroom interaction systems which either have been used in studies relating teacher behavior to pupil outcome or are reasonable representations of the types of systems which have been used in this research. It was the investigators' belief that at least one explanation for the large number of inconsistent and "null" findings in process-product studies was that the instrumentation used to measure classroom behavior, particularly teacher process behavior, may not exhibit convergent and discriminant validity. The findings of this study support this conviction, since about half of the teacher process behaviors investigated failed to pass tests for convergent and discriminant validity. While no reference to specific process-product studies need be made, the investigators suggest that many such studies have measured behaviors with similar forms of instrumentation and some studies have utilized the same instrumentation as was studied in this investigation.

Based upon the results of studying five classroom observation systems, the implications are not particularly encouraging for researchers who choose to measure classroom interaction. One can infer that of the hundreds of other observational coding instruments which have been developed, many must contain categories which do not meet the standards of convergent and discriminant validity proposed in this study. Process-product researchers as well as those who attempt to aggregate and accumulate the findings of process-product research might well be advised to exercise caution in drawing conclusions from studies which use classroom observation systems for which the measurement technique itself accounts for greater variation than the behavior being measured (lack discriminant validity) or that incorporate behaviors which when measured by different systems fail to correlate (lack convergent validity).

As a result of using the MTMM technique to evaluate validity, two types of instrument flaws became apparent. The first concerns the redundancy or overlap of behavioral measures within systems reducing a constructs chances of exhibiting discriminant validity. While complete independence of the behaviors measured within a system is not expected, significant interrelationships among behavioral categories substantially reduce the chances of these categories passing tests for discriminant validity. In several instances in this study, interrelationships among behaviors precluded any chance of a category exhibiting discriminant validity. For example, in the first heterotrait-monomethod triangle in Table 4, CC6 and CC7 correlated .7470, the highest correlation in the matrix. Note that in this instance even if the validity diagonal values had been beyond significance ($r_{.05} = .325$), they probably would not have surpassed the heterotrait-monomethod value and thus the category's discriminant validity would still have been rated "poor". When pilot testing classroom observation instruments, authors might delete highly redundant categories or attempt to reduce the significant interrelationships among such categories by providing more specific operational definitions in order to increase the discriminant validity of their instrument.

The second instrument flaw which came to light with the MTMM technique was the relatively large number (43%) of teacher behaviors which failed to correlate significantly with behaviors on other instruments with which they were matched, i.e., lacked convergent validity. While some of these numbers might be accounted for by the inexactness of the matching process inherent in applying the MTMM technique to classroom observation instruments, in general, the matches that were made in this study may be considered conservative and were often supported by the same or similar operational definitions. Thus, some of the seemingly similar constructs of the type reviewers of process-product studies relate across studies when aggregating process-product findings were, in fact, found not to be similar in this study. One might account for this finding by method variance which confounded the measurement of approximately half the behaviors in this study, vague operational definitions of behaviors when actually interpreted by coders, and intrinsic coder differences. This lack of convergent validity suggests that the descriptive titles of categories and behavioral constructs employed in many observational coding systems may not adequately represent the behavior they purport to measure. Since this flaw is a between-system problem, authors might turn to standard theoretically-based operationalizations of their constructs when constructing new systems.

Evaluating convergent and discriminant validity with the multitrait-multimethod procedure is one approach to assessing the validity of an instrument. The purpose of the remaining portion of this discussion will be to outline both the practical problems encountered in its use in this study and the theoretical assumptions underpinning the technique.

Campbell and Fiske (1959) introduced the technique with examples drawn primarily from the literature in personality and industrial psychology. In these examples, authors attempted to assess various traits (e.g., assertiveness, cheerfulness, poise, popularity, intelligence, etc.) using two or more methods (e.g., self rating vs. peer rating, paper-and-pencil test vs. direct observation, etc.). Thus, the authors of these studies devised different methods for measuring the same variables.

Our use of the technique was somewhat different. Rather than use different methods to measure the same variables, we took existing methods which measured a variety of variables and tried to specify the variables which were measured in common across methods. Thus, our methods and variables were not tailor-made to our research situation. Instead they were fitted to our research needs, and it was this fitting process which created some practical problems.

We found the five systems to be quite different in the way they categorized classroom behavior. One was based upon reinforcement contingencies in the classroom, another on teacher-student interaction, a third sought to categorize teacher behavior so that teacher styles could be described. Each system reflected a particular author's view of the classroom, representing those variables thought to be most important. Thus, different systems sliced the pie of classroom behavior differently, although overlap was apparent. Our approach was to treat the overlap across systems as the basis for constructing comparisons which could be used to determine the convergent and discriminant validity of behavioral categories within systems. However, our success at th's was dependent upon our ability to create fair and accurate matches across systems which defined behavior categories differently.

Often the problem of matching categories reduced to shades of meaning. For example, CC was composed of "giving directions" in Flanders and "gives directions" in CCO. This would seem to be a straightforward match. However, for CC2 we matched "silence or confusion" with "no communication." While perhaps not exactly equivalent, these categories also seemed to have behavior in common. However, matching sometimes became an ambiguous and inexact task. For example, is "telling simple facts" plus "telling complex facts" equal to an "informing statement" (CC21) or do the telling categories include more, or less, than the informing category? Clearly, matching in this case is not the same as in the studies cited by Campbell and Fiske where different methods were designed to measure the exact same variables. The applicability of the MTMM technique to a particular validity problem must ultimately depend on the redundancy of categories and operational definitions across instruments and the conciseness with which matches can be made.

In addition to semantic differences among category definitions, another problem complicated the matching process. This problem involved the differences which sometimes exist betwee the way a category is defined in a manual and the way it is actually used by coders. If a system is to be used reliably by raters, categories must be clearly operationalized for the coders. This, of course, is the purpose of training. However, it is not possible to include in a definition in a manual all of the information necessary to code a particular category reliably. Often coders find it necessary to create "ground rules" to delimit the boundaries of particular categories. Coders, for example, might have difficulty distinguishing between the categories "teacher accepts" and "teacher approves." To distinguish between these behaviors, they may create certain ground rules for coding. For example, coders might decide that if the teacher uses an exclamation such as "Oh!" or "My!" in regard to a student's comment, the proper code is "teacher approves;" otherwise, the code is "teacher accepts." From our experience in this study, ground rules like this are not uncommon, and while they do not seem to distort the meaning of the categories, they delimit their meaning in a way which might not be apparent to a reader of the manual. Furthermore, it is not uncommon for coders or system authors to modify ground rules to fit different classroom situations.

Since the actual operationalizations of categories can change from coder to coder or from study to study (depending upon the classroom situation being coded), the manual definitions, besides being somewhat ambiguous, are at times only guidelines to the meaning of the categories. Thus, the exactitude of the matching process may vary across contexts and coders.

Certain theoretical considerations are also of interest. One of these concerns the independence of methods of measurement. The multitrait-multimethod technique is based upon the use of independent methods of measuring the same variables. Although Campbell and Fiske note that independence is a matter of degree, Calkins, Malitz, Natalicio, and Mote (Note 2), point out that the "determination of validity is enhanced by the inclusion of methods of measurement which are as diverse as possible" (p. 2). The reason for this is that the "determination of convergent and discriminant validity for a set of variables can be obfuscated if all traits have been quantified by the same method of measurement. If all the traits were quantified by the same method, high correlations could result because all the variables share 'method variance'" (pp. 1-2). The extreme case of non-independence is where exactly the same method is used to measure the variables. In this case, the values in the validity diagonal are merely reliability values. Since high reliability can be obtained in the absence of validity, this extreme case would not address the issue of validity. Thus, to the extent that the methods are not independent, the MTMM technique will not yield useful validity data.

In the case of our study, the methods were not as independent as one might wish, since all were based on the use of behavioral observation. While they represented different theoretical rationale and time intervals for collecting data, all could be classified as low to medium inference counting systems producing frequency data. Given that the independence of different classroom observation systems may be difficult to assess, one might include in studies using the MTMM procedure both low inference counting systems and high inference rating scales in order to assure the maximum amount of independence among measurement instruments.

A second theoretical consideration in the use of the MTMM technique involves its statistical assumptions. Kallberg and Kluegel (1975) point out that Campbell and Fiske's technique assumes that traits and methods are uncorrelated and that methods are minimally correlated with one another. Kallberg and Kluegel assert that it is unreasonable to assume that method factors are uncorrelated with trait factors. Moreover, as pointed out above, independence of methods is sometimes difficult to achieve and the degree of independence which exists is difficult to assess. Thus, Kallberg and Kluegel view the model implicit in the MTMM technique as restrictive. As an alternative, they recommend confirmatory factor analysis, CFA (Jöreskog, 1969, 1970), a technique based on the Werts and Linn (1970) path analysis model. Several alternative methods of analyzing MTMM matrices including CFA have been developed, some based upon analysis of variance techniques and some based upon factor analysis (See Alwin, 1970, for a review of these techniques). Of these techniques, it appears that CFA is based upon the least restrictive model (in terms of statistical assumptions about the data). However, CFA cannot be used to its fullest extent unless the matrix contains three or more methods and three or more variables. None of the matrices produced in this study met this requirement. In addition, CFA is based upon rather rigorous statistical and mathematical derivations that are not easily fathomed by the average researcher. This makes the workings of CFA and its output rather difficult to understand and to communicate to other researchers. Until CFA is better understood and more widely accepted, it does not seem to be a practical alternative to Campbell and Fiske's technique, which, despite its assumptions, seems particularly suited to practical psychometric application in process-product research.

This study tested the applicability of the MTMM validation procedure to classroom observation instruments. The study brought to light several nuances and assumptions of the technique which define the context in which MTMM is most appropriate. It was found that the applicability of the MTMM procedure can be expected to vary across validity studies, depending upon two primary considerations: (1) the conciseness in which behavioral categories can be matched across classroom observation systems, and (2) the degree to which the investigator can include comparison instruments in the validity study of sufficient variety, e.g., low vs. high inference, or counting vs. a rating metric, to assure a reasonable degree of independence among methods. To the extent that these considerations are addressed, the validation procedures employed in this study were found to constitute a potentially economical and practical model for examining the validity of other classroom observation systems.

Reference Notes

1. CERLI Verbal-Behavior Classification System (CVC), by Cooperative Educational Research Laboratory, Inc. By permission of Everette Breningmeyer, Cooperative Educational Research Laboratory, Inc., Northfield, Illinois. From a Report to the Office of Education, U. S. Department of Health, Education and Welfare, Contract OEC 3-7-061391-3061, April 1969.

2. Calkins, D., Malitz, D., Natalicio, L., & Mote, T. Validation of some pencil-and-paper anxiety measures by the multitrait-multimethod procedure. Paper presented at the XVI Interamerican Congress of Psychology meeting, Miami, Florida, 1976.

References

Alwin, D. W. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In E. F. Borgatta & G. W. Behmstedt (Eds.), Sociological methodology 1970. San Francisco, California: Jossey-Bass, 1970.

Borich, G. D. Sources of invalidity in measuring classroom behavior. Instructional Science, 6 (3), in press.

Borich, G. D., & Malitz, D. Convergent and discriminant validation of three classroom observation systems: A proposed model. Journal of Educational Psychology, 1975, 67 (3), 426-431.

Brophy, J. E., & Evertson, C. M. Learning from teaching. Boston: Allyn & Bacon, 1976.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Chall, J. S., & Feldman, S. C. A study in depth of first grade reading. (U.S. Office of Education Cooperative Research Project No. 2728). New York: The City College of the City University of New York, 1966.

Flanders, N. A. Analyzing teaching behavior. Reading, Massachusetts: Addison-Wesley, 1971.

Godbout, R. C., Marston, P. T., Borich, G. D., & Vaughan, C. The problem of spurious significance in classroom education research (Res. Rep. 10). Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas at Austin, 1977.

Good, T., & Grouws, D. Teacher rapport: Some stability data. Journal of Educational Psychology, 1975, 67, 179-182.

Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Kallberg, A. L., & Kluegel, J. R. Analysis of the multitrait-method matrix: Some limitations and an alternative. Journal of Applied Psychology, 1975, 60, 1-9.

McDonald, F. J., Elias, P., Stone, M., Wheeler, P., Lambert, N., Calfee, R., Sandoval, J., Ekstrom, R., & Lockheed, M. Final report on phase II beginning teacher evaluation study. Prepared for the California Commission on Teacher Preparation and Licensing, Sacramento, California. Princeton: Educational Testing Service, 1975.

Medley, D. M., & Mitzel, H. E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1959, 49, 227-239.

Rosenshine, B. Teaching behaviours and student achievement. London: International Association for the Evaluation of Educational Achievement, 1971.

Simon, A., & Boyer, E. G. (Eds.) Mirrors for Behavior. Philadelphia: Research for Better Schools, Inc., 1970.

Solomon, D., Bezdek, W. E., & Rosenberg, L. Teaching styles and learning. Chicago: The Center for the Study of Liberal Education of Adults, 1963.

Spaulding, R. L. The Spaulding teacher activity rating schedule (STARS). Durham, North Carolina: Education Improvement Program, Duke University, 1967.

Stallings, J., & Kaskowitz, D. Follow through classroom observation evaluation 1972-1973. Menlo Park, California: Stanford Research Institute, 1974.

Wallen, R. L., Sales, S. M., & Bode, S. Student authoritarianism and teacher authoritarianism as factors in the determination of student performance and attitudes. Journal of Experimental Education, 1970, 38 (4), 83-87.

Werts, C. E., & Linn, R. L. Path analysis: Psychological examples. Psychological Bulletin, 1970, 74, 193-212.

Withal, J., Newell, J. M., & Lewis, W. W. Development of classroom observational categories within a communication model. Madison, Wisconsin: School of Education, University of Wisconsin, 1961.

Footnotes

1. See Borich (1977, Chapter 6) for the results of five large-scale studies which have investigated the relationship between teacher behavior and student outcome and especially pp. 76-78 for a table of consistent and inconsistent findings across these studies.

2. The tendency to (1) report significant findings which fail to exceed the number expected by chance and (2) ignore differences in the operational definitions of purportedly similar constructs serve as examples of the problems which have either reduced the credibility of "significant" process-product findings or led to the proliferation of "null" or inconsistent findings.

Rosenshine's review (1971) illustrates these problems. Rosenshine examined the findings of approximately 50 different studies in which over 200 separate teacher behaviors were investigated. On the basis of evidence from these studies, 11 behaviors were selected as potentially promising in relation to pupil performance. In interpreting the efficacy of these 11 behaviors, however, we must remember that they were derived, for the most part, from correlational, not experimental, studies. Therefore, causation cannot be inferred. Furthermore, these behaviors were derived from clusters of heterogeneous research studies which actually showed mixed results; some studies within a given cluster failed to confirm the efficacy of the variable in question. Also, variables were often operationally defined differently by different investigators. And finally, in some studies the number of significant findings failed to exceed that which could be expected by chance.

The problem of operational definitions is illustrated by the teacher variable <u>clarity</u>, which, Rosenshine points out, has been defined in three very different ways:

(1) whether "the points the teacher made were clear and easy to understand" (Solomon, Bezdek, and Rosenberg, 1963);

(2) whether "the teacher was able to explain concepts clearly . . . had the facility with her material and enough background to answer her children's questions intelligently" (Wallen, 1966);

(3) whether the cognitive level of the teacher's lesson appeared to be "just right most of the time" (Chall and Feldman, 1966).

The problem of chance significance is illustrated by a finding which, I suspect, is not uncommon. Godbout, Marston, Borich, & Vaughan (1977) had occasion to analyze the extent to which process-product relationships in a large-scale teacher effectiveness study replicated over two consecutive years, during which time instrumentation and teacher sample remained constant. Of the 3,050 relationships studied, only 24 were significant at $p < .10$ in the same direction for both years. A much more favorable result would have been expected on the basis of chance alone. Unfortunately, since few replications of this type are conducted, process-product researchers have no way of knowing how unstable their findings may actually be.

Table 1. Coder reliability before and during study.*

System	Highest Prestudy Reliability	Median for 36 Tapes
STARS	72	72
OScAR	88	91
FLANDERS	91	93
CVC	79	82
CCO	86	86

*Scott's coefficient.

Table 2. Flanders vs. CCO

		Flanders		CCO	
		CC1	CC2	CC1	CC2
Flanders	CC1				
	CC2	.6420			
CCO	CC1	.7699	.2779		
	CC2	.5620	.6620	.3454	

Table 3. Flanders vs. OSCAR

		Flanders			Oscar		
		CC3	CC4	CC5	CC3	CC4	CC5
	CC3						
Flanders	CC4	-.2247					
	CC5	-.0684	.1369				
	CC3	.8808	-.1399	.0399			
Oscar	CC4	-.1268	.8571	.2904	-.1331		
	CC5	.2210	-.1297	.0861	.1401	-.0384	

Table 4. CVC vs. STARS

		CVC			STARS		
		CC6	CC7	CC8	CC6	CC7	CC8
	CC6						
CVC	CC7	.7470					
	CC8	.1478	.0072				
	CC6	.0420	.0279	-.2484			
STARS	CC7	.0560	.1766	-.1525	.0121		
	CC8	-.1033	-.1476	.1618	-.1385	0.1730	

Table 5. CCO vs. OSCAR (CCO #18 Mean = 0.00)

		CCO				OSCAR			
		15	16	17	19	15	16	17	19
CCO	15								
	16	-.1576							
	17	-.0101	.1420						
	19	.2797	-.0773	.1237					
OSCAR	15	.4480	.0130	.0973	.0410				
	16	-.3171	.1445	.0030	-.1983	-.0112			
	17	-.2213	.4642	-.0643	.0733	.2283	.1824		
	19	.5054	-.2953	-.1057	.4370	.2738	-.2061	.1677	

Table 6. STARS vs. OSCAR

		STARS		OSCAR	
		20	21	20	21
STARS	20				
	21	-.4358			
OSCAR	20	.6165	-.3520		
	21	-.3314	.8538	-.2478	

Table 7. Flanders vs. CCO vs. OSCAR

		Flanders		CCO		OSCAR	
		1	2	1	2	1	2
Flanders	22						
	23	.2247					
CCO	22	.5743	-.3616				
	23	-.2171	.7782	-.2757			
OSCAR	22	.8808	.1399	.5441	-.1840		
	23	-.1268	.8571	-.2923	.6811	-.1331	

Table 8. CVC vs. OSCAR

		CVC						OSCAR					
		9	10	11	12	13	14	9	10	11	12	13	14
CVC	9												
	10	.0832											
	11	.5459	.1248										
	12	.0858	.5504	.2034									
	13	.1452	.0600	.4429	-.1568								
	14	-.2915	-.0350	-.3095	.0805	.2508							
OSCAR	9	.6746	-.1078	.1829	-.0350	.0058	-.2446						
	10	.3622	.1758	-.0938	.3550	-.0782	-.1198	.1043					
	11	.0236	-.1299	.6402	-.4387	.1818	-.3078	.0494	-.2702				
	12	-.0531	.0866	-.2327	.3088	-.1442	-.1466	-.1845	.2386	-.3163			
	13	.1696	.0200	.3779	-.2166	.6440	.0283	.0684	.0148	.2488	-.2880		
	14	.2008	.2917	-.0933	.3293	.0653	.1928	-.1467	.4183	-.5148	.2790	-.0209	

Table 9. Summary table: CVC vs. OSCAR

comparison category	validity diagonal value	highest value in heteromethod	no. higher than validity value	highest value in monomethod	no higher than validity value
CC9	.6746*	.3622	0	.5459	0
CC10	.1758	.3622	0	.5504	4
CC11	.6402*	-.4387	0	.5459	0
CC12	.3088	-.4387	3	.5504	2
CC13	.6440*	.3779	0	.4429	0
CC14	.1928	.3293	5	-.5148	6

* p < .05

		System I		System II	
		accepts A	questions B	values A	delves B
I	A	(.16)*			
	B	.23	(.70)		
II	A	.43	-.10	(.58)	
	B	-.12	-.01	-.14	(.84)

*Interjudge reliabilities.

Figure 1. Simplified Illustration of the Validation Model. The validity diagonal = .43, -.01; the heterotrait-heteromethod block = .43, -.01, -.10, -.12. The monomethod triangles = .23 and -.14, respectively.

Appendix A

Behavior categories making up each of the comparison categories.

Table	System	General Category Description	Category Numbers in Respective System
1	FLANDERS/CCO		
	CC1	Giving directions	6/7
	CC2	Silence or confusion	10/13
2	FLANDERS/OSCAR		
	CC3	Accepts feeling, praises, encourages	1, 2/2, 12, 22, 32, 42, 52, 62, 72, 82, 92
	CC4	Criticizing or justifying authority	7/6, 16, 26, 36, 46, 56, 66, 76, 86, 96
	CC5	Student talk - initiation	9/10, 20, 30, 40
3	CVC/STARS		
	CC6	Asks for feelings	3/10b
	CC7	Gives feelings	7/10a
	CC8	Disagrees or disapproves	13, 14, 15, 16/1b, 1c, 1d
7	CVC/OSCAR		
	CC9	Informs: facts	5, 6/3, 23
	CC10	Informs: rules	8/4, 5, 7
	CC11	Accepts: facts and interpretations	9, 10/22, 32, 33, 42, 43, 52, 53, 62, 63, 72, 73, 82, 83, 92, 93
	CC12	Accepts: feelings and plans	11, 12/2, 12, 13, 19
	CC13	Rejects: facts and interpretations	13, 14/26, 35, 36, 45, 46, 55, 56, 65, 66, 75, 76, 85, 86, 95, 96
	CC14	Rejects: feelings and rules	15, 16/6, 15, 16, 17

Table	System	General Category Description	Category Numbers in Respective System
4	CCO/OSCAR*		
	CC15	Asks questions	1, 2, 3/8, 50, 60, 70, 80, 90
	CC16	Gives suggestion	6/9
	CC17	Gives direction	7/4, 5, 7, 17, 19
	CC19	Perfunctory agreement/ disagreement	14/14, 34, 44, 54, 64, 74, 84, 94
	STARS/OSCAR		
5	CC20	Restructuring	2b/7
	CC21	Telling, informing	7a, 7b/3
6	FLANDERS/CCO/OSCAR		
	CC22	Accepts feeling, praises, encourages	1, 2/10/2, 12, 22, 32, 42, 52, 62, 72, 82, 92
	CC23	Criticizing or justifying authority	7/12/6, 16, 26, 36, 46, 56, 66, 76, 86, 96

*CCO #18 Mean = 0.00.

Generalizability of Teacher Behaviors Across Classroom Observation Systems

Dick Calkins, Gary D. Borich,
Maria Pascone, and C. L. Kugle
The University of Texas at Austin

Some reviewers of teacher effectiveness research (Borich, 1977a, b; Shavelson & Atwood, 1977; and Shavelson & Dempsey, 1976) have suggested that a possible reason for past failures to find empirically consistent relationships between teacher behavior and pupil achievement is that the measurement process for quantifying both the product and process variables may be unreliable. Since reliability can be viewed as the extent to which a consistent rank ordering can be established among subjects on some variable by a particular measurement procedure, the discovery of potential process-product relationships could be complicated by the inability to accurately rank order teachers according to behavior and students according to achievement by currently existing measurement techniques. In order to investigate the reliability of methods for quantifying teacher behavior, this study obtained data on the generalizability of the behavioral constructs measured by five classroom observation systems. The reliabilities of the items on these classroom observation systems were examined via generalizability theory.

Process-product researchers commonly quantify teacher behavior on a few occasions and then generalize the average score obtained to all other occasions. To the extent that such a score is representative of the behavior for other occasions, the measurement procedure can be said to be reliable. Of the several factors which can potentially cause a behavioral measure to be unreliable, variation due to the conditions under which the observations are made may be the most overlooked.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnum, 1972), which is a combination and extension of classical test theory and Linguist's methods (Linguist, 1953) of multifacet analysis of error, makes possible investigation of the effects of various conditions on the values obtained from and reliability of a behavioral measure. According to generalizability theory any measurement can be thought of as a sample from some large set or universe of measurements of a particular characteristic. Such a universe consists of the set of possible combinations of conditions for which observations could be made. Conditions which vary along the same dimension and under which observations are made are called facets. If the variability introduced by sampling various conditions of a facet is small in magnitude, then a measurement made for any particular condition is representative of the measurement obtainable for any other condition, and hence, the measurement obtained for one condition is generalizable to the entire facet.

In the present investigation of classroom observation systems the facets sampled from the universe of facets were raters and occasions. The particular raters and occasions utilized constitute the conditions or facets of the study. The scores of interest, quantifications of teacher behaviors on various dimensions, were obtained by averaging scores for the available raters and occasions for each teacher. To the extent that the scores for a teacher behavior are comparable across raters and the scores for occasions comparable across occasions, the teaching behavior is considered generalizable. To the extent that scores for a behavioral dimension are generalizable over raters and occasions, rank orderings among teachers on that dimension will be consistent and the likelihood of discovering relationships between pupil achievement and that dimension enhanced.

Decisions about the generalizability of a behavioral dimension as quantified by a particular measurement procedure can be made with the use of a generalizability coefficient which is the counterpart of the reliability coefficient in classical test theory. The coefficient of generalizability which is an interclass correlation (Hays, 1973) is defined as the ratio of the universe score variance to the observed score variance. Like all interclass correlations it takes values between zero and one, and for this coefficient a value of zero indicates total lack of generalizability.

The mechanics of generalizability theory concern analysis of variance components, called facets. These variance components are then used to calculate generalizability coefficients and to suggest changes in a research design, such as increasing the number of raters and occasions needed to obtain a particular level of generalizability. This latter use of the generalizability coefficient is similar to the use made of the Spearman-Brown formula (Nunnally, 1967) in classical test theory. In addition, examination of the variance components of facets can reveal the necessity of making changes in the data collection procedures such as additional training of raters or changes in the overall design such as including more facets, e.g., including a variance component which takes into account differences between subject matter being taught, in order to increase generalizability. The present study undertook to determine the extent to which the behaviors on five classroom observation systems were generalizable across two facets, raters and occasions.

Method

The data were obtained from videotapes of 12 in-service junior high school teachers, each teaching the same content, a unit in social studies, on three occasions of approximately 50 minutes duration. Each of 36 videotapes was rated by five pairs of coders, each pair trained in a different observational coding system. For two of the five observational systems, coders were employed who had previously been trained by the authors of these systems, these being the two most complex systems. The remaining three pairs of coders were trained by the investigators from training materials supplied by system authors and from standard protocols from system manuals.

The five systems employed for this study were selected from Simon and Boyer's Mirrors for Behaviors (1970). The systems were (1) the Observation Schedule and Record, OScAR 5, (Medley & Mitzel, 1959), (2) Spaulding Teacher Activity Rating Schedule, STARS (Spaulding, 1967), (3) Flanders System of Interaction Analysis, (Flanders, 1971), (4) CERLI Verbal-Behavior Classification System, CVC (Cooperative Educational Research Laboratory) (Note 1), and (5) the Classroom Communication Observational System, CCO (Withal, Lewis & Newell, 1961). These systems were selected because of their availability to the educational research community (and therefore presumed use) and for the number of categories and associated operational definitions they had in common.

Upon completion of training, system coders, using their respective systems, rated three trial videotapes of the same general form and content as the experimental tapes in order to obtain estimates of interjudge reliability prior to the study. While reliabilities varied due to system complexity, all were deemed acceptable and are reported in Table 1 along with the median reliability for each pair of coders over all 36 tapes.

Insert Table 1 about here

The data obtained from the five observation systems were analyzed separately utilizing a computer program (Erlich, 1976) which was designed to compute variance components and generalizability coefficients for a fully crossed teacher by occasion by rater design. All facets were assumed to be random.

Results

The results of this study are contained in Tables 2 through 7. Tables 2 through 6 contain the category descriptions and generalizability coefficients for each item on the five observation systems studied. The results for each observation system are presented in a separate table. Table 7 summarizes the generalizability results for comparable categories across systems. In Tables 2 through 7 items are considered generalizable if the generalizability coefficient exceeds .7 for a combination of eight or fewer raters and eight or fewer occasions. Teacher behaviors which require more than these many raters and occasions to obtain a reliable estimate are usually inconsistent and fluctuating, suggesting a need to redefine and/or reconceptualize these variables.

Consideration of Table 2 indicates that six or 43% of the 14 CCO items are generalizable. Consideration of Table 3 indicates that six or 37% of the 16 CVC items are generalizable. Consideration of Table 4 indicates that two or 20% of the ten Flanders items are generalizable. Consideration of Table 5 indicates that 13 or 18% of the 74 OScAR items are generalizable. And, consideration of Table 6 indicates that six or 24% of the 25 STARS items are generalizable. Thus, of the 139 total items in the five classroom observation systems, 33 or 24% were generalizable.

Insert Tables 2 through 7 here

Consideration of Table 7 reveals that of the 11 behavioral categories composed of comparable items across the Flanders, CVC, CCO and STARS systems (no comparable items were found for OScAR) only three categories were generalizable in more than one of these systems. These were the praise and approval category for which the Flanders, CVC and CCO systems had generalizable items; the asks questions category for which the CVC, CCO and STARS systems had generalizable items; and the gives information category for which the Flanders, CVC, CCO and STARS systems had generalizable items.

Discussion and Conclusions

The results of this study indicate that fewer than one-fourth of the behavioral categories on the five classroom observation systems studied were generalizable with any combination of eight or fewer raters and eight or fewer occasions. It is discouraging to note that among process-product researchers these criteria may be considered a particularly liberal standard of generalizability, generally exceeding available resources. In addition, only three behavioral categories logically comparable across systems were generalizable, these behavioral categories being praise and approval, asks questions, and gives information. Moreover, less than half of the behavioral constructs generalizable within systems (15 of 33) were generalizable in any other system, including all combinations of two-system comparisons.

These findings support the contention that either the scores obtained from the five classroom observation systems were not exhibiting those behavioral characteristics of junior high school teachers which are generalizable or much of the teacher behavior as recorded by these classroom observation systems is, in fact, ungeneralizable. In either case if the scores resulting from other classroom observation systems are as unstable as for the five considered in this study, the lack of generalizability may be considered a tenable hypothesis for why so few consistent process-product relationships have been reported.

In addition to the conclusions above, several methodological issues are relevant to the generalizability of the behavioral constructs measured in this study. The implementation of classroom observation systems in process-product research often proceeds by having a sample of teachers observed by a sample of observers on a sample of occasions and their behavior quantified utilizing some observation instrument. Each teacher is then assigned a score for each behavioral category on the observation instrument which is usually the average of the ratings across occasions for that category. Such scores are considered representative of the behavior typically exhibited by that teacher, and thus are used as statistical estimates for purposes of data analysis.

However, a problem arises with the accuracy of this statistical estimate when only the context in which it was calculated is considered instead of all such contexts relevant to the teacher's classroom behavior. The basis of this problem derives from the obvious notion that people behave differently in different situations. The situations or contexts in which teachers behave differently may vary according to the nature of the students, the nature of the school, the subject matter being taught, the resource materials available, past training and experiences, as well as other factors. Thus, situationally determined variation in behavior can cause the determination of generalizability to be misleading when different situations or contexts which affect behavior are typically encountered by the teacher but not considered as facets in a generalizability study. For example, when certain concepts are to be taught to students, one group of teachers because of past training and experience may structure the learning situation such that lecturing is the dominant activity during a particular period (occasion 1) and discussion is the dominant activity during another period (occasion 2). But for a second group of teachers presenting the same content, the first and second periods could both be spent in lecturing and discussion, equaling the amount of time spent in lecturing and discussing by the first group. If a researcher in preparing a generalizability study fails to note such differences among groups of teachers and codes, say, the behaviors <u>gives information</u> and <u>asks questions</u> with periods treated as occasions, neither of these behavioral categories are likely to be found generalizable. For a particular category of behavior to be generalizable from a design which considers raters and occasions as facets, it is necessary that the behaviors coded be consistently emitted and recorded for all occasions. Since only for the second group of teacher will the behavior being coded occur over both occasions, the behaviors will not appear generalizable. Hence, the nature of the teaching situation may mitigate against the generalizability of a particular behavioral category with a design employing only raters and occasions as facets.

An approach to situationally determined variation is to reconceptualize the classroom context by applying generalizability theory to all facets thought to be relevant to the behavioral constructs under study. If "teaching situation" were conceived as a superordinate categorization of teacher behavior, such as giving information, asking questions, providing reinforcement, etc., then each teaching behavior within such categorizations could be characterized by a score for each situation. Such situation specific scores might be more likely found generalizable over raters and occasions and related to student achievement than overly simplistic, context-free behavioral constructs as presently conceptualized by some classroom observation systems.

Reference Note

1. CERLI Verbal-Behavior Classification System (CVC), by Cooperative Educational Research Laboratory, Inc. By permission of Everette Breningmeyer, Cooperative Educational Research Laboratory, Inc., Northfield, Illinois. From a Report to the Office of Education, U. S. Department of Health, Education and Welfare, Contract OEC 3-7-061391-3061, April 1969.

References

Borich, G. D. The appraisal of teaching: Concepts and process. Reading, Massachusetts: Addison-Wesley, 1977a.

Borich, G. D. Sources of invalidity in measuring classroom behavior. Instructional Science, 1977b, 6 (3).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measures: The theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc., 1972.

Erlich, O. A study of the generalizability of measures of teacher behavior. Unpublished dissertation, University of California, Los Angeles, 1976.

Flanders, N. A. Analyzing teaching behavior. Reading, Massachusetts: Addison-Wesley, 1971.

Hays, W. L. Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton-Mifflin, 1953.

Medley, D. M., & Mitzel, H. E. A technique for measuring classroom behavior. Journal of Educational Psychology, 1959, 49, 227-239.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Shavelson, R., & Atwood, N. Generalizability of measures of teaching process. In Borich, G. D. The appraisal of teaching: Concepts and process. Reading, Massachusetts: Addison-Wesley, 1977.

Shavelson, R., & Dempsey, N. Generalizability of measures of teaching behavior. Review of Educational Research, 1976, 46 (4), 553-611.

Simon, A., & Boyer, E. G. (Eds.) Mirrors for behavior. Philadelphia: Research for Better Schools, Inc., 1970.

Spaulding, R. L. The Spaulding teacher activity rating schedule (STARS). Durham, North Carolina: Education Improvement Program, Duke University, 1967.

Withal, J., Newell, J. M., & Lewis, W. W. Development of classroom observational categories within a communicatioh model. Madison, Wisconsin: School of Education, University of Wisconsin, 1961.

Table 1. Coder reliability before and during study.*

System	Highest Prestudy Reliability	Median for 36 Tapes
STARS	72	72
OScAR	88	91
FLANDERS	91	93
CVC	79	82
CCO	86	86

*Scott's coefficient.

Table 2

Item descriptions and generalizability coefficients for two raters and three occasions for CCO.

Item	Title	Generalizability Coefficient
1	Asks Information	.1001
2	Seeks or accepts direction	.5839*
3	Asks for opinion or analysis	.4601
4	Listens	.2498
5	Gives information	.6403*
6	Gives suggestion	.6958*
7	Gives direction	.4281
8	Gives opinion	.6610*
9	Gives analysis	.6956*
10	Shows positive feeling	.7220
11	Inhibits communication	0 NR
12	Shows negative feeling	.3621
13	No communication	0 NV
14	Prefunctory agreement or disagreement	.4646

Note: Coefficients above .7 without an (*) are generalizable with two raters and three occasions

* Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions

NR represents the situation when no responses were coded for this item

NV represents the situation when a negative variance less than -.2 occurred for this item

Table 3

Item descriptions and generalizability coefficients for two raters and three occasions for CVC.

Item	Title	Generalizability Coefficient
1	Seeks factual or specific information	.2027
2	Asks for reasoning, explanation, interpretation, judgment, or evaluation	.3133
3	Asks for feelings, asks about feelings	.6658*
4	Asks about rules, plans, or directions	.4463
5	Tells factual specific material	0 NV
6	Gives reasons, interpretations, judgments or evaluation	.7188
7	Gives or tells feelings	.4093
8	Tells classroom structure, rules, directions, plans	.2835
9	Approves factual or specific answers	.6101*
10	Accepts reasoned ideas, interpretations, judgments	.3464
11	Approves or empathizes with feelings expressed	.1020
12	Accepts or agrees with plans, rules, directions	.4160
13	Disagree with answer or factual statement	.7760
14	Disapproves of thinking, interpretation, etc.	.5708*
15	Responds negatively to feelings expressed by others	0
16	Rejects rules, plans, expectations, directions	.5437*

Note: Coefficients above .7 without an (*) are generalizable with two raters and three occasions.

* Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions

NR represents the situation when no responses were coded for this item

NV represents the situation when a negative variance less than -.2 occurred for this item

Table 4

Item descriptions and generalizability coefficients for two raters and three occasions for Flanders.

Item	Title	Generalizability Coefficient
1	Accepts feeling	.0357
2	Praises or encourages	.7686
3	Accepts or uses ideas of student	.1690
4	Asks questions	0 NV
5	Lecturing	.8593
6	Giving direction	0 NV
7	Criticizing or justifying authority	.0974
8	Student talk-response	0 NV
9	Student talk-initiation	.4764
10	Silence or confusion	0 NV

Note: Coefficients above .7 without an (*) are generalizable with two raters and three occasions.

* Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions

NR represents the situation when no responses were coded for this item

NV represents the situation when a negative variance less than -.2 occurred for this item

Table 5

Item descriptions and generalizability coefficients for two raters and three occasions for OScAR.

Item	Title	Generalizability Coefficient
1	Considering statement	.3881
2	Informing statement	.5865*
3	Describing statement	0 NV
4	Directing statement	.0268
5	Rebuking statement	0
6	Desisting statement	.2496
7	Non-substantive question	.0250
8	Procedural question/positive	.7133
9	Pupil non-substantive/utterance	0
10	Pupil non-substantive/not evaluated	0
11	Pupil non-substantive/supported	.6652*
12	Pupil non-substantive/approved	.7293
13	Pupil non-substantive/acknowledged	0
14	Pupil non-substantive/neutrally rejected	0
15	Pupil non-substantive/criticized	0
16	Pupil procedural question/negative	.4977*
17	Pupil procedural question/neutral	.0978
18	Pupil procedural question/positive	0
19	Pupil question	0
20	Pupil question/not evaluated	0
21	Pupil question/supported	0
22	Pupil question/approved	.0038
23	Pupil question/acknowledged	0
24	Pupil question/neutrally rejected	0
25	Pupil question/criticized	0 NR
26	Pupil statement	0
27	Pupil statement/not evaluated	.0198
28	Pupil statement/supported	.8549
29	Pupil statement/approved	.0119
30	Pupil statement/acknowledged	0

Table 5 (cont.)

Item	Title	Generalizability Coefficient
31	Pupil statement/neutrally rejected	.1896
32	Pupil statement/criticized	0 NR
33	Pupil response	0
34	Pupil response/not evaluated	0 NV
35	Pupil response/supported	.6623*
36	Pupil response/approved	.3718
37	Pupil response/acknowledged	.2019
38	Pupil response/neutrally rejected	0
39	Pupil response/criticized	0 NR
40	Problem structuring/statement	.3071
41	Choral response	.5257*
42	Choral response/supported	0
43	Choral response/approved	.2034
44	Choral response/acknowledged	.4573
45	Choral response/neutrally rejected	.6751*
46	Choral response/criticized	0 NR
47	Convergent question/not answered	.0431
48	Convergent interchange/not evaluated	.2081
49	Convergent interchange/supported	0
50	Convergent interchange/approved	0
51	Convergent interchange/acknowledged	0
52	Convergent interchange/neutrally rejected	.0433
53	Convergent interchange/criticized	0 NR
54	Elaborating question/not answered (1)	.1562
55	Elaborating interchange/not evaluated (1)	0
56	Elaborating interchange/supported (1)	.2479
57	Elaborating interchange/approved (1)	0
58	Elaborating interchange/acknowledged (1)	.5074*
59	Elaborating interchange/neutrally rejected (1)	0
60	Elaborating interchange/criticized (1)	0 NR
61	Elaborating question/not answered (2)	.5519*
62	Elaborating interchange/not evaluated (2)	0
63	Elaborating interchange/supported (2)	.5499*
64	Elaborating interchange/approved (2)	.4514

Table 5 (cont.)

Item	Title	Generalizability Coefficient
65	Elaborating interchange/acknowledged (2)	.1201
66	Elaborating interchange/neutrally rejected (2)	0
67	Elaborating interchange/criticized (2)	0 NR
68	Divergent question/not answered	0
69	Divergent interchange/not evaluated	.3014
70	Divergent interchange/supported	.5769*
71	Divergent interchange/approved	.0914
72	Divergent interchange/acknowledged	0
73	Divergent interchange/neutrally rejected	0
74	Divergent interchange/criticized	0 NR

Note: Coefficients above .7 without an (*) are generalizable with two raters and three occasions.

* Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions

NR represents the situation when no responses were coded for this item

NV represents the situation when a negative variance less than -.2 occurred for this time

Table 6

Item descriptions and generalizability coefficients for two raters and three occasions for STARS.

Item	Title	Generalizability Coefficient
1	Non-transactional behavior	.0
2	Disapproval with aversive stimuli present	0 NR
3.	Disapproval indicated by removal of social reinforcers or, in some cases, physical reinforces	0 NR
4	Withholding reinforcers when a student or child bids for attention	.1917
5	Approval with positive affect present	.4103
6	Social and/or motor structuring	.2204
7	Social and/or motor restructuring	.0755
8	Digressions	.6173*
9	Inductive methods-presenting simple facts	0 NR
10	Inductive methods-complex concepts	0
11	Concept formation-simple facts (deductive)	0
12	Concept formation-complex facts (deductive)	.5782*
13	Concept formation-simple or complex events (transductive or analogical)	0 NR
14	Telling-simple facts	.7106
15	Telling-complex facts	0 NR
16	Rote process-simple	0
17	Information	.4719
18	Focusing attention	.3616
19	Asking for or eliciting recall-simple	.5247*
20	Asking for or eliciting recall-complex	0
21	Asking for use or application-simple	.1968
22	Asking for use or application-complex	.1991
23	Expressing-values, opinion, feelings	.7221
24	Eliciting student expressions-values, opinions, feelings	.3010
25	Listening to or observing non-teached directed pupil activity	.6382*

Note: Coefficients above .7 without an (*) are generalizable with two raters and three occasions.

*Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions

NR represents the situation when no responses were coded for this item

NV represents the situation when a negative variance less than -.2 occurred for this item

Table 7

Generalizability coefficients for two raters and three occasions for logically comparable items for the Flanders, CVC, CCO, and STARS observation systems.

Behavioral Category	Item	Flanders	Item	CVC	Item	CCO	Item	STARS
Approval of feelings	1	.0357	11	.1020	10	.7220	5	.4103
Praise and approval	2	.7686	9	.6101*	10	.7220	5	.4103
			10	.3464				
			11	.1020				
			12	.4160				
Accepts or uses student's ideas	3	.1690	12	.4160	2	.5839*	5	.4103
Asks questions	4	0	1	.2027	1	.1001	8	.6173*
			2	.3133	2	.5839*	9	0
			3	.6658*	3	.4601	10	0
			4	.4463			19	.5247*
							20	0
							21	.1968
							22	.1991
							24	.3010
Gives information	5	.8593	5	0	5	.6403*	11	0
			6	.7188	8	.6610*	12	.5782
			7	.4093	9	.6956*	13	0
							14	.7106
							15	0
							16	0
							17	.4719
							23	.7221

Table 7 (cont.)

Behavioral Category	Item	Flanders	Item	CVC	Item	CCO	Item	STARS
Gives direction	6	0	8	.2835	7	.4281	6	.2204
							7	.0755
Disapproval	7	.0974	14	.5708*	12	.3621	2	0
			15	0			3	0
			16	.5437*				
No communication	10	0			13	0	1	0
							25	.6382*
Informational feedback-negative			13	.7760			3	0
Ignoring					11	0	4	.1917
Listens/observes					4	.2498	25	.6382*

Note: Coefficients above .7 without (*) are generalizable with two raters and three occasions.

[*] Generalizability of .7 was reached for some combination of 8 or fewer raters and 8 or fewer occasions.

Measuring Classroom Interactions: How Many Occasions are Required to Measure Them Reliably?

Oded Erlich
Tel-Aviv University
Israel

and

Gary Borich
The University of Texas
at Austin

One important line of inquiry in research on teaching has operationally defined teacher behavior variables and examined their relationships to student achievement. This approach assumes that the individual teacher plays a key role in producing student learning. However, results from correlational studies of teacher behaviors and student outcomes have been disappointing with most correlations low or nonreplicable (Borich, 1977; Shavelson and Atwood, 1977).

Shavelson and Atwood (1977) in a recent review of current research on teaching, hypothesized that one possible reason for the lack of relationships between teacher behaviors and student achievement is that the generalizability of behavioral measurements has not been adequately examined or established to allow conclusions about relationships between teacher behavior and student outcomes. Measures of teacher behavior contain potential sources of error (facets) such as observation occasion, observers, and subject matter which might affect their generalizability. The generalizability of measures is not a function of these separate facets, but rather a function of the simultaneous consideration of all the facets which might affect the generalizability of the measures. The effect of these facets on the generalizability of teacher behavior can be estimated by the application of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972). In generalizability theory a generalizability study (G study) is conducted which has two purposes. The first is to examine the generalizability of teacher behavior measures by considering the measurement facets (e.g., occasions and raters) which affect the reliability of measurements obtained. Based on this analysis, a G study then recommends variables for inclusion in future decision studies (D studies) which examine relationships between teacher behaviors and student outcomes.

Only a few studies on the generalizability of teacher behavior measures have been reported. They have either explained how to apply generalizability theory to examine the problems in measuring teacher behavior (e.g., Medley and Metzel, 1963; McGaw, Wardrop, and Bunda, 1972; Marzano, 1973; and Rowley, 1975), or they have failed to use appropriately the data available (e.g., Sandoval, 1974).

Since generalizability studies were not available, Shavelson and Atwood explored the measurement problem by examining studies which differed in measurement facets, but which used similar teacher variables. They attempted to find patterns in the stability of teacher behavior measures across these studies. Although they could not determine the amount of error contributed by various facets in separate studies, they reached several tentative conclusions about the stability of 13 clusters of teacher behavior variables. Global ratings were found to be the most stable measures while variable clusters of teacher presentation, positive feedback, probing, and direct control were found to be moderately stable. Unstable variable clusters included teacher questions, negative feedback, nonprobing behaviors, indirect control, and student-centered teaching style.

A recent generalizability study (G study) examined patterns of error sources contributed by the facets of raters and occasions in order to identify generalizable measures of teacher behavior (Erlich, 1976). Erlich analyzed data collected by Sandoval (1974) which included frequency counts as well as global ratings on five 5th grade teachers observed by two or three raters while teaching reading and mathematics on three observation occasions. Erlich defined a measure of a variable to be generalizable if it required a combination of 4 or less raters and 10 or less occasions to reach a coefficient of generalizability of 0.7. Variables were classified into three groups: (1) low frequency of occurrence (ultimately excluded from analysis), (2) high frequency variables whose measures appeared not to be generalizable, and (3) high frequency variables whose measures appeared to be generalizable. The teacher behavior variables found to be potentially generalizable in his analysis supported most of the conclusions reached by Shavelson and Atwood concerning the stability of these variables. The one exception occurred in the cluster of variables related to teacher questions. Shavelson and Atwood found all questioning behaviors to be unstable, but Erlich found that those questioning variables related to ways of checking student reactions and learning were generalizable.

The purpose of the present study was to examine the generalizability of measures of classroom interaction occurring during 2nd and 3rd grade class activities excluding reading instruction and to provide information concerning the number of observation occasions required to reach a 0.7 level of generalizability for each of these measures.

Method

The data analyzed in this study were collected during the second year of a two year replicated study of teacher effectiveness using the Brophy-Good Teacher-Child Dyadic Interaction System (Brophy and Evertson, 1976). Subjects were 28 teachers who had 5 or more years of teaching experience with their 3 most recent years of experience at the 2nd or 3rd grade level. These teachers were selected because they had shown high consistency in producing student learning gains on the Metropolitan Achievement Tests. They were observed between 9-14 times during whole class activities and reading instruction by two different raters who alternated across occasions.

Some teachers were eliminated from our analysis since almost no data were recorded for them on one or both of the two main categories of variables in the Brophy-Good System--public response and private response variables--during nonreading class activities. Also, although teachers were observed a total of 9-14 times, about half the occasions included only reading group instruction, leaving 3-7 occasions in which data were collected on dyadic interactions other than reading. Therefore, the sample was reduced and data analyzed consisted of 17 teachers on 5 occasions for the category of public response variables and 22 teachers on 5 occasions for the category of private response variables.

The design selected for the analysis was a one facet nested design; occasions being nested within teachers. Occasions were considered to be nested because teachers were observed at different times of day, on different days and teaching what may be considered different lessons.

Even though an implicit source of error, raters were not considered as a potential source of error in this analysis for several reasons. First, all raters had extensive training during the first year of the study and during the summer prior to the second year of the study, enabling them to consistently reach at least 0.8 agreement. Furthermore, the criteria for agreement required that raters achieve a 0.8 reliability not only for their codes in each general category of behavior in the observation system, but also for frequency counts on clusters of variables within each category. Disagreements between raters were most often a result of one rater being able to code more information than another, and, therefore, the rank ordering of the teachers was not affected. This implies that there was minimal teacher-rater interaction; and therefore, raters were considered not to be a potential source of error affecting the generalizability of the measures.

The Teacher-Child Dyadic Interaction observation instrument attempts to code all dyadic interactions (teacher behaviors with respect to an individual child as well as the child's response and interactions with the teacher) occurring in the classroom. It contains 167 variables divided into two main categories: public response variables, in which the teacher-child interaction occurs in a group setting; and private response variables, in which the teacher and child confer privately about the child's individual work. Within these two categories of variables, Brophy and Good identified clusters of variables. The public variables included the following clusters: Teacher's Method of Selecting Students to Respond; Difficulty Level of Questions; Type of Questions Asked (Academic or Nonacademic); Quality of Student Response to Questions; Teacher's Feedback Reaction to Student Responses; Student Initiated Comments; and Student Initiated Questions. The private interaction variables were divided into three clusters: Child Created Contacts (CCC); Teacher Afforded Contacts (TAC); and Behavior Related Contacts.

Generalizability theory was used as the statistical basis for the data analysis. For each variable, the analysis provided the estimate of the universe score (true score in classical test theory) variance [$\hat{\sigma}^2(t)$], and the estimate of the error variance, which in this design was due to the teacher occasion interaction confounded with the occasion variance and unidentified sources of error [$\hat{\sigma}^2(o,to,e)$]. Based on these variance components, the number of occasions required to obtain a generalizability coefficient of 0.7 was calculated for each variable. A generalizable variable was defined in this study as one for which a coefficient of generalizability of 0.7 could be obtained by observing the teacher on ten or fewer observation occasions. Not only is ten a practical upper limit on the number of observation occasions which could be used, but also, and of greater importance, teacher behaviors which require more than ten occasions to obtain a reliable estimate are usually inconsistent and fluctuating, suggesting a need to redefine and/or reconceptualize these variables.

Results

Initial inspection of the data revealed that a majority of the variables occurred infrequently. Two types of low frequency variables were identified. The first type of infrequent variable consisted of variables for which the frequencies were scattered throughout the teacher by occasion matrix, i.e., only a small number of teachers engaged in these dyadic interactions, the frequency counts for these interactions were uniformly low and were obtained on less than three occasions for each teacher. This type of infrequent variable was eliminated from the analysis since these frequency counts were too inconsistent, too low, limited to too few teachers, and would have demanded a very large number of occasions to reach an acceptable level of generalizability. Two entire clusters of variables--Student-Initiated Questions and Student-Initiated Comments--and one sub-cluster--Opinion Questions--were completely eliminated by this process. Brophy and Evertson (1976) suggested in their analysis that these types of interactions may be inappropriate for teaching fundamental tool skills in the 2nd and 3rd grade. The other low frequency variables of this type which were also eliminated were spread throughout the remaining variable clusters. In general, these variables appeared to be infrequent because of the detailed nature of the observation instrument which attempts to allow for all possible interactions even when their occurrence is not probable (e.g., praise after a wrong answer or criticism after a right answer).

A second type of low frequency variable was retained for analysis. These low frequency variables were recorded for relatively few teachers, but occurred more consistently. These variables, although occurring infrequently, may be generalizable, and, if such is the case, should be included in correlational studies of teacher behaviors and student outcomes.

Table 1 presents the results of the analysis for the public response variables. Variables are grouped into four clusters based on those developed by Brophy and Good. Each variable cluster is discussed separately. For each variable the table includes the estimates of universe score variance [$\hat{\sigma}^2(t)$] and error variance [$\hat{\sigma}^2(o,to,e)$] and the number of occasions required to reach a 0.7 level of generalizability.

INSERT TABLE 1 ABOUT HERE

The first variable cluster, Teacher's Selection of Respondents, describes the way in which the teacher selects students to respond to questions asked. The teacher may either preselect (name the child who is to answer before asking the question), select a child from among those who volunteer to answer, or select a nonvolunteer. If a student gives the answer before the teacher has time to select a student, this is labeled a "call-out." Relatively few occasions, three, are needed to obtain a reliable measure of the frequency of call-outs. Teacher "selection of a volunteer" and the "preselection of a student" to respond are generalizable, but these variables require more occasions, five and eight respectively, to reach a 0.7 level of generalizability. The last variable, "selection of a nonvolunteer," requires twelve occasions and is nongeneralizable, if we use our earlier criterion of ten occasions as the practical upper limit of the number of occasions that are possible.

The next cluster, Type of Questions, contains variables related to the type of questions asked. "Product" and "process" questions represent difficulty levels of academic questions. To answer a product question, the child must give a specific correct answer which can be expressed in a single word or short phrase. The process question, which is more complex, requires the child to explain the steps which must be followed to solve a problem or reach a conclusion. "Math questions" do not differentiate between the difficulty level of the questions, but include all questions related to math content. The last two question variables, although both subject-matter related, are considered nonacademic questions. These are called self-reference questions because they are not intended to elicit a particular correct factual answer, but ask the child instead about some factor in his personal background.

Three variables in this cluster were found to be generalizable. "Math questions" and "subject-matter-related self-reference questions about the student's experience" can both be estimated by the use of three occasions. "Product questions," the type found to occur most frequently at this grade level, require six occasions. The two remaining question variables, "process questions" and "subject-matter-related self-reference questions asking a student's preference," are nongeneralizable, requiring 17 and 48 occasions, respectively.

The third cluster, Quality of Student Response to Questions, evaluates student answers to questions. Four variables were considered: "correct" and "part-correct," "wrong," and "no response." The number of "correct" and "wrong" answers can be estimated by using six occasions, and the number of "no responses" by using ten occasions. However, measurement of the variable "part-correct responses," requires twelve occasions, and is therefore nongeneralizable.

The last cluster involves public response opportunities and contains variables of Teacher Feedback Reaction to Student Responses. Three types of feedback which occur after a correct student response: "praise," "process feedback" (explaining the process involved in reaching the correct answer), and "asking a new question" are generalizable, requiring two, two, and four occasions respectively. Teacher feedback which "affirms the answer" following a correct response is nongeneralizable, requiring the use of 12 observation occasions. The three remaining types of feedback occurred after either a wrong answer or a no response. All are generalizable with "asking another student" requiring the use of five occasions, "rephrasing or cluing after a wrong answer" requiring eight occasions, and "asks another student" requiring four occasions.

Table 2 presents the results of the analysis for the private dyadic interaction variables. In these interactions, the teacher deals privately with one child about matters idiosyncratic to the child. These interactions may be "work-related" (giving or asking for help with class content or procedures), "personal" (giving or asking for personal information or for favors), or "behavior-related" (classroom behavior). The personal and work-related interactions are divided into two clusters: "Child Created Contacts" (CCC) and "Teacher Afforded Contacts" (TAC). The "Behavior-Related Contacts" (which are all teacher afforded) are clustered separately to correspond to the Brophy-Good classification.

INSERT TABLE 2 ABOUT HERE

The first variable cluster, "Child Created Contacts," contains more variables than any other cluster. Not only does there appear to be many child created dyadic interactions at the 2nd and 3rd grade level, but most of these interactions are generalizable. Only three variables are nongeneralizable: "content-related CCC" (12 occasions), "content-related CCC giving long feedback" (24 occasions), and "work-related procedural CCC which were delayed" (15 occasions). Eight of the remaining eleven variables can be estimated reliably by the use of four or fewer occasions, suggesting that many child created contacts are highly consistent behaviors at the 2nd and 3rd grade levels.

The second variable cluster, "Teacher Afforded Contacts" (TAC) is composed of private interactions initiated by the teacher. Most teacher afforded contacts are work related. The total TAC "math contacts" can be estimated reliably by the use of three observation occasions, while TAC "work contacts with long feedback" and TAC "work contacts with brief feedback" require 5 and 7 occasions respectively to reach a 0.7 coefficient of generalizability. TAC "work contacts involving criticism" are also generalizable (6 occasions). However, TAC "work contacts involving observation" require the use of 17 occasions and are nongeneralizable. The other teacher afforded private interactions in this cluster involved "procedural management contacts" or "personal contacts." The former required the use of only 3 occasions and was generalizable, while the latter was nongeneralizable, needing 12 occasions.

The last cluster, Behavior Relatrd Contacts, contains 5 variables. The first 3 describe types of teacher reactions to student misbehavior. "Teacher criticism" can be estimated reliably by the use of 3 occasions, while "teacher warnings" and "nonverbal intervention" require 5 and 9 occasions respectively. The last 2 variables assess the teacher's handling of behavior-related contacts. Contacts in which "no teacher error" occurs can be estimated reliably by the use of only 2 occasions, and contacts involving "teacher overreaction" require 7 occasions.

In summary, the findings indicate that many public and private variables can be considered as generalizable if measured by the required number of observation occasions. On the other hand, many other variables obtained such low frequency counts that they were excluded from analysis and are considered to be nongeneralizable. The large number of infrequent teacher-child dyadic interaction variables leaves open the possibility that dyadic interactions may consist of a more limited range of behaviors than originally conceptualized, at least at the primary level. Classroom observations at higher grade levels might still show that some of the infrequent variables eliminated from this analysis do occur more frequently and/or consistently at these levels. If such is the case, these variables should be analyzed to determine their generalizability within the framework of these higher grade levels.

Conclusions

This study shows that the generalizability of behavioral measurements must be an important consideration in classroom observation research. This analysis has identified generalizable measures of classroom interactions at the 2nd and 3rd grade level and determined the number of observation occasions required to reach a 0.7 level of generalizability. It should be recalled that raters were not considered as an error source in this study because extensive training of raters and the stringent criteria for *a priori* inter-rater agreement ensured a high inter-rater reliability and a minimum of teacher-rater interaction. When observers are trained to an appropriate level as in the Texas Teacher Effectiveness Study (Brophy and Evertson, 1976), they may be eliminated as a source of error. Otherwise, raters as well as occasions must be considered as potential sources of error affecting the generalizability of the measures.

Past classroom observation studies have often used three or fewer observation occasions to measure teacher behaviors. This study found that many variables must be measured by more than three occasions to obtain a 0.7 level of generalizability. Future studies collecting classroom observational data in the lower grade levels using the Brophy-Good system or similar systems measuring these same behaviors should rely upon the findings of thic study and use the number of observation occasions required to estimate the measures reliably.

It is apparent that generalizable classroom measures require different numbers of occasions to be measured reliably. Therefore, it seems appropriate to recommend that the generalizability of different measures obtained by other observation systems be examined in order to determine how to measure them reliably. Obtaining reliable measurements will enable researchers to eliminate sources of measurement error which may be contributing to the lack of relationships between classroom interactions and student outcomes.

Table 1

Estimate of Universe Score Variance and Error Variance, and Number of Occasions Required to Reach 0.7 Level of Generalizability for Public Dyadic Interaction Variables

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
TEACHERS'S SELECTION OF RESPONDENTS			
Call-outs by student	4.64	5.83	3
Selects Volunteer	65.21	127.03	5
Preselects student	14.13	66.73	8
Selects Nonvolunteer	73.07	237.96	12
TYPE OF QUESTION			
Academic Questions			
Total math response opportunities	175.23	241.59	3
Product questions	135.62	373.89	6
Process questions	0.54	11.03	48
Nonacademic Questions			
Subject-matter-related self-reference question about the student's experience	6.08	7.37	3
Subject-matter-related self-reference question asking student's preference	2.27	16.94	17
QUALITY OF STUDENT RESPONSE TO QUESTIONS			
Correct	117.94	289.69	6
Wrong	4.87	12.99	6
No Response	3.84	14.96	9
Part-correct	0.79	3.72	12

Table 1 (continued)

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
TEACHER FEEDBACK REACTION TO STUDENT RESPONSES			
Following Correct Answer			
Process feedback	1.07	0.81	2
Praise	16.72	16.50	2
Asks new question	11.70	22.73	4
Affirms answer	5.04	24.72	12
Following Wrong Answer			
Asks another student	1.31	2.63	5
Rephrases or clues	0.36	1.26	8
Following No Response			
Asks another student	2.23	4.24	4

Table 2

Estimate of Universe Score Variance and Error Variance, and Number of Occasions Required to Reach 0.7 Level of Generalizability for Private Dyadic Interaction Variables

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
CHILD CREATED CONTACTS			
Work-Related Interaction			
Content			
Total Math child created work contacts	69.68	143.76	5
Content-related CCC given brief feedback	5.09	21.42	10
Content-related CCC	9.12	48.37	12
Content-related CCC given long feedback	1.28	12.94	24
Procedural			
Work-related procedural CCC	13.57	15.45	2
Work-related procedural CCC with Criticism	0.09	0.14	4
Work-related procedural CCC with Praise	0.04	0.12	8
Work-related procedural CCC which were delayed	0.05	0.29	15
Work-related procedural CCC with brief feedback	15.33	11.27	2
Personal Interactions			
CCC personal experience sharing interactions	2.38	4.75	4
CCC personal procedural interactions	6.41	12.08	4
CCC personal procedural interactions which were granted	2.04	6.31	7
CCC personal procedural interactions not granted	0.98	1.38	3

Table 2 (continued)

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
TEACHER AFFORDED CONTACTS			
Total Math teacher afforded work contacts	21.05	28.44	3
Work contact with long feedback	9.17	20.11	5
Work contact with brief feedback	28.01	83.92	7
Work contact involving observation	1.09	8.07	17
Work contact involving criticism	0.97	2.43	6
Procedural management contacts	124.28	137.75	3
Personal contacts	0.92	4.63	12
BEHAVIOR RELATED CONTACTS			
Teacher criticism	4.40	5.57	3
Teacher warnings	16.77	27.94	4
Nonverbal intervention	0.52	2.07	9
No teacher error	45.32	43.38	2
Teacher overreaction	0.22	0.70	7

References

Borich, G. D. Sources of invalidity in measuring classroom behavior. Instructional Science, 1977, in press.

Brophy, J. E., & Evertson, C. M. Learning from teaching: A developmental perspective. Boston: Allyn and Bacon, Inc., 1976.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: The theory of generalizability for scores and profiles. New York: John Wiley and Sons, Inc., 1972.

Erlich, O. A study of the generalizability of measures of teacher behavior. Unpublished doctoral dissertation, University of California, 1976.

Marzano, W. A. Determining the reliability of the Distar instructional system observation instrument. Unpublished thesis, University of Illinois, 1973.

McGaw, B., Wardrop, J. L., & Bunda, M. A. Classroom observation schemes: Where are the errors? American Education Research Journal, 1972, 9(1), 13-27.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage, Handbook of research on teaching. Chicago: Rand McNally, 1963.

Rowley, G. L. The reliabilities of classroom observational measures: Estimation, interpretation, application. Unpublished doctoral dissertation, University of Toronto, 1975.

Sandoval, J. Beginning teacher evaluation study completion report, task K: Sub-study of consistency of teacher behavior. Princeton. New Jersey and Berkeley, California: Educational Testing Service, 1974.

Shavelson, R., & Atwood, N. Generalizability of measures of teaching process. In G. D. Borich, The appraisal of teaching: Concepts and process. Reading, Massachusetts: Addison-Wesley, 1977.

Generalizability of Teacher Process Behaviors during Reading Instruction

Oded Erlich
Tel-Aviv University
Israel

and

Gary Borich
The University of Texas
at Austin

Attempts to find correlations between reading instruction and reading achievement have previously centered around methods of teaching reading (e.g., whole word vs. phonics) (Chall, 1967). While some tentative conclusions have been drawn about the relative effectiveness of various methods, no one method has been shown to be unquestionably superior. One important approach for studying factors related to reading achievement is that of observing operationally defined variables of teacher behavior and classroom interaction and then relating them to reading achievement. This approach assumes that pupil-teacher classroom interactions play a key role in producing pupil learning. By identifying classroom interactions which increase pupil achievement, researchers can assist teachers in constructing an empirically validated instructional model for the teaching of reading.

Results from past correlational studies of teacher behaviors and student outcomes (including, but not restricted to reading achievement) have been disappointing, with most correlations low or nonreplicable (Shavelson and Atwood, 1977). One possible reason for the lack of relationship between classroom interactions and student achievement is that the generalizability of behavioral measurements has not been adequately examined or established to allow conclusions about relationships between teacher behavior and student outcomes to be drawn. In this paper we will be concerned with the generalizability of classroom interaction measures during reading instruction.

The Concept of Generalizability

The concept of generalizability is based on the notion that the behavior observed represents only a sample of the true behavior. If the sample of observed measurements contain little or no error, the generalization to the characteristic (true) behavior is sound; the accuracy of the measurement is high. If the observed scores contain sizable error of measurement, the generalization to the characteristic behavior is tenuous; the accuracy is low. Measures of teacher-pupil classroom interaction contain potential sources of error (facets) such as observation occasion, observers, subject matter, etc. Only by considering the effect of all these facets can we determine the extent to which teacher behavior measures are generalizable.

For example, in most studies of teaching process, a random sample of teachers is observed by two or more raters. The consistency with which the teachers are rank ordered on some variable such as "number of verbal reinforcements" or "number of questions the teacher asks" is interpreted as the reliability of the measurement. Typically each teacher's score is an average of the raters' scores for that teacher and is usually interpreted as characteristic of the teacher asking questions or using verbal reinforcements. No doubt that the use of several raters provides a more precise measure on each teacher but what about the nature of the pupils taught, the teaching situation, the subject-matter taught, and other factors that might contribute to the instability of the teachers' behavior? While the measurement is taken in one particular setting and at one particular point in time, it is usually interpreted as generalizing over many settings at different points in time.

Only a few studies on the generalizability of teacher behavior measures have reported on more than one facet. Most have either explained how to apply generalizability theory to examine the problems in measuring teacher process variables or they have failed to use appropriately the data available. (See Erlich, 1976.) Two appropriate generalizability studies recently examined variables of student-teacher classroom interaction. Erlich and Borich (1976) analyzed classroom interactions during nonreading class activities in the 2nd and 3rd grades. Erlich (1976) analyzed 5th grade teacher behaviors occurring during reading and math combined. Because different subject matters, e.g., reading, math, social studies, may elicit different kinds and frequencies of pupil-teacher classroom interactions, observation data of interactions occurring during different subject matters may need to be examined separately.

Purpose

The purpose of this study was to identify teacher-pupil interactions occurring during beginning reading instruction and to examine the generalizability of these measures of classroom interaction.

Method

Sample. The data analyzed in this study were collected during the second year of a two year replicated study of teacher effectiveness using the Brophy-Good Teacher-Child Dyadic Interaction System (Brophy and Evertson, 1976). Subjects were 26 teachers who had 5 or more years of teaching experience with their 3 most recent years of experience at the 2nd or 3rd grade level. These teachers were selected because they had produced consistent pupil learning on the Metropolitan Achievement Tests over three consecutive years.[1] Teachers were observed from between three and seven times during teachers' reading instruction by two different raters who alternated across occasions. Four

[1] A linear pattern of either gain, constancy, or decline over the three-year period constituted the definition of consistent pupil learning in this study (Brophy, 1973).

teachers who had been observed on less than five occasions were eliminated from our analysis. For those teachers who were observed on more than five occasions, five occasions were selected at random for the analysis. Thus, the final data analyzed included 22 teachers each observed on five occasions.

Design. The design selected for the analysis was a one facet nested design; occasions being nested within teachers. Occasions were considered to be nested because teachers were observed at different times of day, on different days and teaching what may be considered different lessons.

Even though an implicit source of error, raters were not considered as a potential source of error in this analysis for several reasons. First, all raters had extensive training during the first year of the study and during the summer prior to the second year of the study, enabling them to consistently reach a 0.8 agreement. Furthermore, the criteria for agreement requirement that raters achieve the 0.8 reliability not only in their coding for each category in the observation system, but also on frequency counts within each category. Disagreements between raters were most often a result of one rater being able to code more information than another, and, therefore, the rank ordering of the teachers was not affected. This implies that there was also a minimal teacher-rater-interaction; and therefore, raters were considered not to be a potential source of error affecting the generalizability of the measures.

Instrument. The instrument used to collect data was the Teacher-Child Dyadic Interaction Observation System (Brophy and Good, 1969). This instrument attempts to code all dyadic interactions (teacher behaviors with respect to an individual child as well as the child's response and interactions with the teacher) occurring in the classroom. It contains 167 variables divided into two main categories: public response variables, in which the teacher-child interaction occurs in a group setting; and private response variables, in which the teacher and child confer privately about the child's individual work.

Within these two categories of variables, Brophy and Good identified clusters of variables. The public variables included the following clusters: Teacher's Method of Selecting Students to Respond; Difficulty Level of Questions; Type of Questions Asked (Academic or Nonacademic); Quality of Student Response to Questions; Teacher's Feedback Reaction to Student Responses; Student Initiated Comments; and Student Initiated Questions. The private interaction variables were divided into three clusters: Child Created Contacts (CCC); Teacher Afforded Contacts (TAC); and Behavior Related Contacts.

Statistical Analysis. The effect of the occasion facet on the generalizability of teacher-child interactions was estimated by the application of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnum, 1972). In generalizability theory a generalizability study (G study) has two purposes. The first is to examine the generalizability of the measures (e.g., of teacher behavior) by considering the potential sources of error (e.g., occasions and raters) which affect the reliability of measurements obtained. Based on this analysis, a G study then recommends variables for inclusion in future decision studies (D studies) which examine, for example, relationships between teacher behaviors and student outcomes.

For each variable examined in this study, the G study analysis provided the estimate of the universe score (true score in classical theory) variance [$\hat{\sigma}^2(t)$], and the estimate of the error variance, which in this design was due to the teacher occasion interaction confounded with the occasion variance and unidentified sources of error [$\hat{\sigma}^2(o,to,e)$]. The formula for obtaining the coefficient of generalizability in this design is $\hat{\rho}^2 = \dfrac{\sigma^2(t)}{\sigma^2(t) + \sigma^2(o,to,e)/n}$ where n is the number of occasions. Using this formula and based on the estimates of the variance components, the number of occasions (n) required to obtain a prespecified level of generalizability can be calculated for each variable.

A generalizable variable was defined in this study as one for which a coefficient of generalizability of 0.7 could be obtained by observing the teacher on ten or fewer observation occasions. Not only is ten a practical upper limit on the number of observation occasions which could be used, but also, and of greater importance, teacher behaviors which require more than ten occasions to obtain a reliable estimate are usually inconsistent and fluctuating, suggesting a need to redefine and/or reconceptualize these variables.

Results

Initial inspection of the data revealed that a majority of the variables occurred infrequently, inconsistently, and were recorded for only a few teachers. This pattern of occurrence was characteristic of all variables in three clusters--Student-Initiated Questions, Student-Initiated Comments, and Child Created Contacts--and two sub-clusters--Opinion Questions and Non-Academic Self Reference Questions. Brophy and Evertson (1976) suggested in their analysis that the classroom interactions represented by these variables may not be appropriate for teaching fundamental tool skills such as reading and math in the 2nd and 3rd grades. The rest of the low frequency variables were scattered throughout the remaining variable clusters. They appeared to be infrequent mainly because of the detailed nature of the observation instrument which attempts to allow for all possible interactions even when their occurrence is not likely (e.g., praise after a wrong answer or criticism after a right answer). None of the low frequency variables described above appeared to play any appreciable role in primary reading instruction in the classrooms observed and were, therefore, eliminated from the generalizability analysis.

Another type of low frequency variable was retained for analysis. These variables differed from those previously described in that the behaviors occurred for at least 20% of the teachers. These variables may be important in distinguishing between effective and ineffective teachers despite their relatively infrequent occurrence across teachers and their generalizability should be examined. Those found to be generalizable should be included in correlational studies of teacher-pupil classroom interaction and student outcomes to determine if they are, in fact, important variables in reading instruction.

Table 1 presents the results of the analysis for the classroom interaction variables analyzed. Variables are grouped into five clusters based on those developed by Brophy and Good (1969). The first four clusters contain public interactions, and the last cluster contains private interactions. Each variable cluster is discussed separately. For each variable the table includes the estimates of universe score variance [$\hat{\sigma}^2(t)$] and error variance [$\hat{\sigma}^2(o,to,e)$] and the number of occasions required to reach a 0.7 level of generalizability.

INSERT TABLE 1 ABOUT HERE

The first variable cluster, Teacher's Selection of Respondents, describes the way in which the teacher selects students to respond to questions asked. The teacher may either preselect (name the child who is to answer before asking the question), select a child from among those who volunteer to answer, or select a nonvolunteer. If a student gives the answer before the teacher has time to select a student, this is labeled a "call-out." Relatively few occasions are needed to obtain a reliable (generalizable) measure of the selection of a volunteer, or a non-volunteer or of the frequency of call-outs (2, 3, and 4 respectively). The last variable, "preselection of a student" is also generalizable, but requires more occasions (9) to reach a 0.7 level of generalizability.

The next cluster, Type of Question, contains variables related to the type of questions asked. "Choice questions," "product questions," and "process questions" represent difficulty levels of academic questions. To answer a choice question, the child must select the correct answer from two or more options given by the teacher. To answer a product question, the child must give a specific correct answer which can be expressed in a single word or short phrase. The process question, which is the most complex, requires the child to explain the steps which must be followed to solve a problem or to reach a conclusion. Two of the three variables in this cluster were found to be generalizable. "Product questions" and "choice questions," the types found to occur most frequently in reading instruction at these grade levels, require four and five occasions respectively to reach a 0.7 level of generalizability. "Process questions" is nongeneralizable, requiring 16 occasions to reach the acceptable level of generalizability.

The third cluster, Quality of Student Response to Questions, evaluates student answers to questions. Four variables were considered: "correct" and "part-correct," "wrong," and "no response." All can be estimated by three or fewer occasions, indicating that of these variables the behaviors are highly consistent within a particular reading instruction group.

Only one variable in the Teacher Feedback Reaction to Student Responses cluster--praise following a correct answer--occurred frequently enough to warrant analysis. Apparently, this is the only type of feedback which occurs regularly during reading instruction. It needs only three observation occasions to obtain a 0.7 level of generalizability.

The last cluster, Teacher Afforded Contacts (TAC) contains private dyadic interactions. TACs may be related to work, to procedures, or to a child's behavior. Only a few variables in this cluster were analyzed because most behaviors occurred infrequently. The measures of TAC variables related to work and to management procedures were both nongeneralizable. These teachers' behaviors, although occurring frequently, fluctuated so greatly that 13 and 18 occasions would be needed to obtain a reliable estimate of their behavior. On the other hand, measures of interactions related to a child's behavior were quite consistent. All measures of behavior-related contacts are generalizable with the number of occasions required to reach a 0.7 level of generalizability ranging from 3 to 5.

Discussion

The findings above indicate that a majority of the variables analyzed can be considered as generalizable if measured by the required number of observation occasions. It should be recalled, however, that all other Dyadic Interaction System variables not presented in the table exhibited such low frequency counts that they were excluded from analysis. Although some of these might be found generalizable, this generalizability statistically could result from the fact that their frequency of occurrence tends to be consistently zero.

The large number of infrequent teacher-child dyadic interaction variables suggests that primary reading instruction consists of a limited range of such behaviors. These findings, however, do not exclude the possibility that some classroom interaction variables during reading instruction at higher grade levels might be more infrequent and/c consistent at these levels. If such is the case, these variables should be analyzed to determine their generalizability.

Ten observation occasions were selected as the maximum number allowed to reach a 0.7 level of generalizabiilty in this study. The number of occasions required to reach this level for those variables which were generalizable ranged from 1-9 occasions. Past classroom observation studies considering a range of subject matters and grade levels, have often used three or fewer occasions to measure teacher behaviors (Shavelson and Atwood, 1977). The present analysis indicates that some variables require more than three occasions to be measured reliably. It should be noted, however, that in this study interactions occurring frequently during reading instruction may, in general, be considered highly consistent. Almost half of the generalizable variables could be measured reliably by the use of three observation occasions and approximately three quarters of them by the use of five observation occasions.

Classroom observation studies frequently observed teachers teaching different subject matters, but combined different subject matters for analysis. The Teacher Effectiveness Study (Brophy and Evertson, 1976) coded the reading data separately, allowing reading and non-reading class activities to be analyzed separately. A comparison of the results of this study with those of Erlich and Borich (1976), who analyzed the generalizability of the non-reading activities, indicates that classroom interactions during reading and non-reading instruction differ in several significant ways.

Reading instruction appears to be primarily a public process. With the exception of behavior-related contacts, almost all of the private interaction variables occurred infrequently. Non-reading class activities appeared balanced between public and private interactions and included many more private teacher-child interactions (both teacher afforded and child created). For example, in Erlich and Borich's analysis, the cluster of child created contacts contained the largest number of variables analyzed. In this study, the entire cluster was eliminated because so few instances of child created contacts during reading instruction were recorded.

Teachers also asked different types of questions in reading and non-reading instruction. During non-reading activities, almost all questions asked were "product questions." "Choice questions" appeared so infrequently that this variable was not even analyzed. During reading instruction, however, choice questions occurred frequently and were highly generalizable (four occasions).

Teachers appeared to find choice questions particularly suited to reading instruction, but not to other subjects. Teacher questions were more task oriented during reading instruction. Self-reference questions were asked during non-reading activities, but only academic questions occurred during reading instruction.

Teacher behaviors appeared influenced by the reading context in several other important ways. For example, selection of a nonvolunteer during non-reading activities was inconsistent and its measurement nongeneralizable, while the same behavior was highly consistent and its measurement generalizable during reading instruction. The more consistent selection of nonvolunteers during reading suggests that the teacher is more likely to insist upon involving the reluctant, shy, or non-assertive child during reading than during non-reading activities. Another noteworthy difference occurred in the quality of student responses to questions. The percentage of correct, wrong, part-correct, and no-response answers could be estimated in three or fewer occasions during reading instruction, while the number of occasions required during non-reading activities was six or greater. This difference suggests that the teacher is more consistent in gauging the difficulty level of questions during reading instruction than during other activities. A final difference was that feedback type reactions were far more limited during reading instruction than during non-reading instruction. Only one feedback response--praise after a correct response--was employed frequently enough during reading instruction to be considered for analysis.

In summary, the findings of this study suggest that observation data for reading instruction should be analyzed separately from data obtained during other types of instruction. Behaviors observed during, say, math or social studies may not occur during reading, and conversely, reading instruction may elicit behaviors unique to that context. This study found that reading instruction encompassed a narrower range of pupil-teacher classroom interaction than that found during non-reading instruction in the same classrooms. Even when the same behaviors occurred across subject matters, measures of these behaviors may be generalizable in one context and not in the other; or the number of occasions necessary to reach an acceptable level of generalizability may differ. In planning future observational studies of reading instruction, researchers should rely upon the findings of this study to ascertain the appropriate number of observations needed to obtain generalizable measures of teaching behavior during reading instruction.

Table 1

Estimate of Universe Score Variance and Error Variance, and Number of Occasions Required to Reach 0.7 Level of Generalizability for Dyadic Interaction Variables during Reading Instruction

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
Teachers' Selection of Respondents			
Selects volunteer	105.83	161.93	2
Selects Nonvolunteer	258.09	381.72	3
Call-outs by student	10.86	19.49	4
Preselects student	14.68	59.74	9
Type of Question			
Choice questions	162.45	266.64	4
Product questions	273.78	608.93	5
Process questions	2.42	16.64	16
Quality of Student Response to Questions			
Part-correct	5.69	3.15	1
Correct	384.24	342.09	2
Wrong	19.09	21.09	3
No Response	6.96	10.43	3
Teacher Feedback Reaction to Student Responses			
Praise following correct answer	35.34	41.34	3

(Table continued on next page.)

Table 1 (cont.)

	$\hat{\sigma}^2(t)$	$\hat{\sigma}^2(o,to,e)$	Number of Occasions
Teacher Afforded Contacts			
Work contact involving brief contact	5.41	30.45	13
Procedural management contacts	5.55	42.80	18
Behavioral related contacts.			
Contacts involving no teacher error	8.45	11.08	3
Contacts involving teacher warning	4.80	7.87	4
Contacts involving teacher criticism	0.97	2.22	5

References

Brophy, J. E. Stability in teacher effectiveness. American Educational Research Journal, 1973, 10, 245-252.

Brophy, J. E., & Evertson, C. E. Learning from teaching: A developmental perspective. Boston: Allyn and Bacon, Inc., 1976.

Brophy, J. E., & Good, T. Teacher-child dyadic interaction: A manual for coding classroom behavior. Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1969.

Chall, G. Learning to read: The great debate. New York: McGraw-Hill Book Company, 1967.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: The theory of generalizability for scores and profiles. New York: John Wiley and Sons, Inc., 1972.

Erlich, O. A study of the generalizability of measures of teacher behavior. Unpublished doctoral dissertation, University of California, 1976.

Erlich, O., & Borich, G. D. Measuring classroom interactions: How many occasions are required to measure them reliably? Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1976.

Shavelson, R., & Atwood, N. Generalizability of measures of teaching process. In G. D. Borich, The appraisal of teaching: Concepts and process. Reading, Massachusetts: Addison-Wesley, 1977.