DOCUMENT RESUME

ED 142 586 TM 006 427

AUTHOR Erlich, Oded; Borich, Gary
TITLE Generalizability of Teacher Process Behaviors during Reading Instruction.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
PUB DATE 15 Oct 76
NOTE 20p.; For related documents, see ED 042 688, 066 438, and 127 333

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Data Analysis; Elementary School Students; Feedback; Grade 2; Grade 3; *Interaction Process Analysis; Participant Involvement; Predictor Variables; Primary Education; Reading; *Reading Achievement; *Reading Instruction; Statistical Analysis; *Student Teacher Relationship; *Teacher Behavior; Test Reliability
IDENTIFIERS *Generalizability Theory; Teacher Child Dyadic Interaction System (Brophy)

ABSTRACT

Maintaining that the generalizability of behavioral measures has not been sufficiently established to permit conclusions about the relationship between teacher behavior and student achievement, the present research examines the generalizability of classroom interaction variables measured by the Brophy-Good Teacher-Child Dyadic Interaction System during 2nd and 3rd grade reading instruction. Using generalizability theory as the statistical basis for data analysis, the number of measurement occasions required to reach the 0.7 level of generalizability for five clusters of classroom interaction variables were identified. Analyses revealed that the interaction characterizing reading instruction differs from that characterizing other kinds of instruction in regard to: 1) proportion of public to private teacher-pupil interactions; 2) nature of questions asked; and 3) teacher behavior concerning feedback, pupil involvement, and question difficulty. (Author)

The *Evaluation of Teaching Project*—one of four projects at the Research and Development Center for Teacher Education—has as its mission to develop materials and strategies for teacher training, research and evaluation. The goals of the *Evaluation of Teaching Project* are to develop (1) a conceptual framework for the evaluation of teaching, (2) a sourcebook of validated teacher evaluation and research instruments, and (3) strategies for the evaluation of teacher trainee programs. These goals are being carried out with funds from the National Institute of Education.

In the process of meeting its objectives, the *Evaluation of Teaching Project* conducts systematic research in teacher behavior for the purpose of validating instruments and identifying characteristics of effective teaching. The following report describes one facet of this research. A complete listing of studies in this report series is available by writing to the Evaluation of Teaching Project, R&D Center for Teacher Education, University of Texas at Austin, 78712.

Generalizability of Teacher Process Behaviors during Reading Instruction

Oded Erlich
Tel-Aviv University
Israel

Gary Borich
The University of Texas
at Austin

October 15, 1976

Abstract

Maintaining that the generalizability of behavioral measures has not been sufficiently established to permit conclusions about the relationship between teacher behavior and student achievement, the present research examines the generalizability of classroom interaction variables measured by the Brophy-Good Teacher-Child Dyadic Interaction System during 2nd and 3rd grade reading instruction. Using generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972) as the statistical basis for data analysis, the number of measurement occasions required to reach the 0.7 level of generalizability for five clusters of classroom interaction variables were identified. Analyses revealed that the interaction characterizing reading instruction differs from that characterizing other kinds of instruction in regard to: 1) proportion of public to private teacher-pupil interactions; 2) nature of questions asked; and 3) teacher behavior concerning feedback, pupil involvement, and question difficulty.

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 <u>generalizing</u> 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 $\hat{\rho}^2 = \frac{\sigma^2(t)}{\sigma^2(t) + \sigma^2(o,to,e)/n}$ 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/or 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.