DOCUMENT RESUME

ED 049 272 TM 000 444

AUTHOR Maxey, James H.

TITLE Analysis of Observational Data.

PUB DATE Feb 71

NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29

DESCRIPTORS Analysis of Variance, *Classroom Observation Techniques, *Data Analysis, Educational Research, Evaluation Methods, Measurement Techniques, *Reliability, *Research Design, *Research Methodology, Research Problems, Sampling, Statistical Analysis, Teacher Behavior, Teacher Evaluation, Validity

ABSTRACT

Three common methodological problems in the analysis of observational data include: (1) failure to base degrees of freedom for the analysis on the appropriate sampling unit; (2) the validity of the design and data are inconsistent with the generalizations reached; (3) the statistic used to calculate coder reliability is often inappropriate and yields values that are misleading. A simulated study on the effects of observational training on teacher performance is employed to illustrate these problems. Hypotheses, research procedure, observer reliability, results, data analysis, and conclusions are included in the simulated report, which is criticized in terms of the problems outlined. (AE)

ERIC Full Text Provided by ERIC

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED049272

ANALYSIS OF OBSERVATIONAL DATA

by

JAMES H. MAXEY
GEORGIA STATE UNIVERSITY

As a doctoral candidate, my learned professors muttered such things as appropriate sampling unit, questionable reliability and validity when discussing research based on observational data. Being a good student, I solemnly nodded my head in agreement. Now that I have passed that magical threshold of knowledge, I go around muttering similar things; however, my students are so rude as to ask me what these strange mutterings mean. So, in order to clarify my own thinking and to be able to share these ideas with my students, I developed the following paper which hopefully provides clear illustrations of three common problems in the analysis of observational data. Three of the main difficulties include: (1) Researchers base the degrees of freedom for the analysis on the number of students rather than the number of teachers, even though the teachers represent the sampling unit and the researcher wishes to generalize the results to teacher training or behavior. (2) The validity of the design and data are inconsistent with the generalizations reached. (3) The statistic used to calculate coder reliability is often inappropriate and yields values that are misleading.

My students (through the use of student category 13, "Praise of Professor" and category 14, "Student Use of Professor Ideas") have indicated that the article was helpful. I hope that you find it helpful for your staff and students.

In order to facilitate the explanation of these problems, a simulated research report follows. Examples from the research report will be used to demonstrate the three common problems. This report is abbreviated to in-

TM 000 444

clude only the sections necessary for illustration of clear examples and is not intended to be a complete report.

The Effects of Observational Training on Teacher Performance: A Simulated Report

Since observational training is being used in teacher preparation programs, it is important to know if this training affects the verbal performance of teachers in the classroom and if the changes affect attitudes of the students about the teacher. It is reasonable to assume that the successful teacher has a wide repertoire of verbal behaviors. Moreover, given this repertoire, the teacher will be more likely to choose an appropriate teaching style for any given occasion. The current research defines this concept as teacher flexibility. Several studies indicate that teacher flexibility is related to positive student attitudes. A different set of studies indicate that inservice workshops in observational training increase flexibility of teachers. The study reported in this paper is a partial replication of these current studies.

Hypotheses

The two major research hypotheses are:

H_1: Teachers participating in the inservice observational workshop will be more flexible in their teaching style than those teachers not participating in the workshop.

H_2: Teachers participating in the workshop will receive higher student ratings than those teachers not participating in the workshop.

H_2: form two

Students taught by teachers with higher flexibility ratios will have more positive attitudes than students taught by teachers with lower

ERIC Full Text Provided by ERIC

flexibility ratios.

Procedure

The sample consisted of forty tenth grade mathematics teachers from a large urban area. The teachers all volunteered to participate in a five day inservice workshop that was held during Thanksgiving vacation. The workshop consisted of learning to code verbal behavior, participating in skill practice exercises and role playing teaching situations.

Two hours of observational data was collected on each teacher, and a student attitude questionnaire was administered four weeks prior to the inservice training workshop. In order to insure that the experimental and control groups were initially equivalent, the subjects were matched on a flexibility ratio and on student attitude ratings. This blocking arrangement was used to randomly assign the subjects to either the treatment or the control group.

The instrument used to collect data was Flander's basic ten category system. Flexibility was defined as cells:

$$\frac{(4,3) + (5,3) + (8,3) + (3,4) + (3,5) + (4,9) + (9,4) + (9,5)}{\text{Total Tallies}}$$

This definition was chosen because it is an estimate of how often for short periods of time a teacher shifts from direct to indirect patterns of influence.

After the training sessions two additional hours of observational data was collected on each teacher, and the student attitude questionnaire was administered again. Although a two-way, fixed model, analysis

ERIC Full Text Provided by ERIC

of variance with repeated measure on one factor would be most appropriate, a post-test ANOVA was used in this report for illustraion purposes.

Observer Reliability

The reliability coefficient used to check the observer was calculated by the Scott Index using the following formula

$$\text{Scott Index} = \frac{P_o - P_e}{100 - P_e}$$

where P_o was the percent agreement calculated by subtracting the percent error between two observers from 100. P_e was found by squaring the percentage of tallies falling into each category, dividing each product by 100, and then summed over all categories.

An estimate of inter-coder reliability was based on a two session sample. Both observers coded the same session during the first week of observations and again after the training session. The reliability for the first session was .73 and for the second .78. This was judged to be an adequate level of reliability.

Results

Teacher flexibility referred to ability of the teacher to switch from one style presentation to another within a given lesson. The cell means are presented in Table 1, while the ANOVA is illustrated in Table 2.

TABIE 1

CELL MEANS FOR TEACHER FLEXIBILITY

	Pre - Test	Post - Test
Experimental Group	.0816	.1132
Control Group	.0824	.0899

TABLE 2

ANOVA FOR FLEXIBILITY

Source of Variation	D.F.	Adjusted Sum of Squares	Mean Square	F
Between	1	1.42	1.42	4.24**
Within	38	12.54	.33	
Total	39			

The difference between the experimental group and the control group was statistically significant at the .01 level.

"Student attitude" as measured by the student questionnaire referred to a positive regard for the teacher and classroom. The cell means for student attitudes are presented in Table 3, and the ANOVA in Table 4.

TABLE 3

CELL MEANS FOR STUDENT ATTITUDE

	Pre-Test	Post-Test
Experimental Group	168.73	173.41
Control Group	169.21	170.32

TABLE 4

ANOVA FOR STUDENT ATTITUDE

Source of Variation	D.F.	Adjusted Sum of Squares	Mean Square	F
Between	1	3254.88	3254.88	5.26**
Within	998	617552.42	618.79	
Total	999	620807.30		

The difference between the experimental group and the control group was statistically significant at the .01 level.

Conclusions

The first hypothesis which predicted that trained teachers would be more flexible was supported by data. The second hypothesis which predicted that the trained teachers would receive higher student rating was supported by the data. The educational implications are that observational training is useful in helping teachers to have a wider repertcire of teaching behaviors and that students find these new styles of teaching satisfying. Therefore, observational training has much to offer as an inservice training technique.

Criticism of Simulated Study

The first major criticism relates to the analysis of the second hypothesis. The degrees of freedom used in the ANOVA was based on the total number of students (1000) which is clearly inappropriate. The proper degrees of freedom should have been based on the number of teachers (40). There are two major reasons for this observation as follows:

(1) The teachers were the basis for drawing the sample.

(2) The purpose of the study was made to make generalizations about certain teaching behaviors and the effects of observational training of teachers.

One of the reasons for this type of error is that the researcher lets the data collection process determine the analysis. In other words, since the researcher has 1000 student questionnaires, he uses them for the analysis. If the second hypothesis would have been analyzed using class means for the student ratings and the number of teachers for the degrees of freedom the results would not have been significant. See Table 5.

TABLE 5

PROPER ANOVA FOR HYPOTHESIS 2

Source of Variation	D.F.	Adjusted Sum of Squares	Mean Square	F
Between	1	1243.40	1243.40	1.20
Within	38	39374.46	1036.17	
Total	39	40617.86		

Hypothesis 2 written in form 2 raises an interesting question. It appears as if the hypothesis is generalizing to student behavior and that the number of students would be the proper sampling unit. It is my opinion that both hypotheses have exactly the same purpose; however, form 2 is more subtle. The same argument would apply to the analysis of the data. Teachers constitute the sampling unit and the proper degrees of freedom.

A second major criticism relates to questions of validity. The rationale implies that flexibility is the ability of the teacher to pick an appropriate teaching style for a given situation. Yet, the definition refers to how often a teacher switches from direct behavior to indirect behavior within a given lesson. This represents a weak tie between the concept and the measurement. The measurement does not guarantee that the switching represents desirable behavior. A better way would be to define desirable teaching behaviors on some theoretical basis. Researchers have been unwilling or unable to define desirable behavior.

Some studies define flexibility as the more cells used in matrix the more flexible the teacher. For example, the teacher who has tallies in 60 cells out of a 100 cell matrix is defined as more flexible than the teacher who has tallies in 40 cells. In part, the same weaknesses exist in this example as in the above definition. The definition is not tied to any theoretical construct of appropriate behavior. It may indicate that more variety on the part of teachers exists when observations are made over a longer period of time.

A third common way of defining flexibility is to give the teacher two lesson plans, one which is highly structured and one which is unstructured. The teacher is then observed teaching both lessons, and flexibility is defined as the difference between the two I/D ratios. The I/D ratio in Flander's category system is based on ratio of indirect behaviors (such as "empathy", "praise","use of student ideas", and "questions") to direct behaviors (such as "lecturing", "criticizing", and "giving directions"). This method probably comes the closest to a reasonable definition of flexibility. It is still too global an approach to really pinpoint appropriate behaviors.

One of the basic problems with all three methods is that they take into account only two chain sequences of events. It may well be that in order to look for appropriate behaviors longer chains will have to be recorded. It is possible that present recording methods which are all based on a two-chain approach are insufficient to capture the important differences.

A second question concerning validity relates to the theoretical relationship between the inservice workshop and increased student ratings. Even if this relationship was consistently true, it does not provide sufficient data for planned change. This raises several questions. What does the teacher do differently that pleases the students? Which changes are most important in changing student attitude? In general this relationship is a good example of the "mystic of education". It does not build a strong rationale for the relationship of the treatment to the results.

A third concern relates to the contamination effect between treatment and measurement. During the workshop the teachers are taught to code teaching behaviors using a specific coding system while the observations for measuring teacher change are based on the same or similar coding techniques. It seems logical that there are some unknown contaminations between the treatment and the measurement. For example, some of the changes may be superficial and merely represent the teacher acting to please the code. It would be desirable for the researchers to find additional ways to measure the effects of such training.

A third major criticism of the study is with regards to reliability. Reliability in this study is based on category totals, and yet, the subsequent analysis is based on a flexibility ratio consisting of certain cell totals. Since it is quite possible to have high category agreement

ERIC Full Text Provided by ERIC

and low cell agreement in matrix coding, this method is clearly an unsatisfactory estimate of inter-code reliability. The true reliability regarding flexibility ratio is unknown. It may be considerably higher or lower th an the coefficient listed.

In general, this inadequacy is representative of the failure to relate the calculation of inter-code reliability to the particular type of analysis to be performed. This would suggest that there is a need to calculate several reliabilities, one for each different type of analysis used. Data analysis is often based on the sequence cell totals or category totals. This is not considered in the calculation of the inter-coder reliability.

Matrix coding represents both transitional cells (changing from one category to another) which are psychologically determinate and steady state cells (time spent in a single category) which are indeterminate. The transition cells are determinate since it is a matter of direct observation whether a particular transition occurs or not. Tallies in the cteady state cells are indeterminate. They depend on the size of the time interval used. Due to the combination of two different types of data, it is difficult to base reliability on a combination of the two. The reliability of transitional events is probably of primary interest. New methods must be developed for the calculation of inter-coder reliability that allow for the comparison between codes of either the total or any given part of a transitional matrix. The fact that there is a growing interest in multi-chain coding (the recording of transitions that include sequences of events of a larger length than two) complicates the problem even more.