REPORT RESUMES

ED 011 520 SF 000 968

THE RELIABILITY OF OBSERVATIONS OF TEACHERS' CLASSROOM BEHAVIOR.

BY- BROWN, BOB BURTON AND OTHERS

PUB DATE 67

EDRS PRICE MF-$0.09 HC-$1.08 27P.

DESCRIPTORS- BEHAVIOR RATING SCALES, EDUCATIONAL PHILOSOPHY, FILMS, *LESSON OBSERVATION CRITERIA, PERSONAL VALUES, *RELIABILITY, *STATISTICAL ANALYSIS, TABLES (DATA), TEACHER BACKGROUND, *TEACHER CHARACTERISTICS, TEACHER EVALUATION, TEACHER PRACTICES OBSERVATION RECORD, GAINESVILLE

THIS PORTION OF AN "INVESTIGATION OF OBSERVER-JUDGE RATINGS OF TEACHER COMPETENCE" WAS PRIMARILY DEVOTED TO STATISTICAL ISSUES IN ASSESSING THE RELIABILITY OF OBSERVATIONS OF TEACHERS' CLASSROOM BEHAVIOR. FROM 67 TO 130 STUDENT TEACHING SUPERVISORS, ACADEMIC PROFESSORS, AND EDUCATION PROFESSORS FROM TWO LARGE MIDWESTERN UNIVERSITIES AND TWO "TEACHERS COLLEGE-TYPE" INSTITUTIONS VIEWED FIVE 10-MINUTE FILMS OF CLASSROOM TEACHING ON ONE OCCASION AND TWO OF THE FIVE FILMS AGAIN A YEAR LATER. AFTER EACH OF THE TWO VIEWINGS, SUBJECTS RESPONDED TO THE 62-ITEM TEACHER PRACTICES OBSERVATION RECORD, ON WHICH THE OBSERVER CHECKED THOSE OF THE LISTED PRACTICES OBSERVED DURING THE FILM VIEWING. RELIABILITY FINDINGS WERE THAT CORRELATIONS (1) OF OBSERVERS' TOTAL SCORES WITHIN A GIVEN FILM VIEWING WERE VERY GOOD, (2) OF OBSERVERS' TOTAL SCORES BETWEEN REPEAT FILM VIEWINGS ONE YEAR APART WERE POOR TO FAIR, (3) BETWEEN-OBSERVER RELIABILITY WERE FAIR, (4) WITHIN-OBSERVER RELIABILITY WERE FAIR, AND (5) OF INTERNAL CONSISTENCY RELIABILITY WERE VERY GOOD. (LC)

THE RELIABILITY OF OBSERVATIONS OF TEACHERS' CLASSROOM BEHAVIOR

by

Bob Burton Brown, Director
Teacher Competence Research Project, W1 Building K
College of Education
University of Florida, Gainesville

William Mendenhall, Chairman
Department of Statistics
University of Florida, Gainesville

Robert Beaver, Instructor
Department of Statistics
University of Florida, Gainesville

The Teacher Practices Observation Record (TPOR) is an instrument for measuring classroom behavior by systematic observation. It attempts to measure the agreement-disagreement of teachers' observed classroom behavior with educational practices advocated by John Dewey in his philosophy of experimentalism. In addition to presenting this instrument and briefly describing its development, we will report the reliability data obtained by using it in a study of observations of filmed teaching episodes. The data reported on the TPOR will be placed in the context of the general problem of studying reliability, and will be used to demonstrate a new design for estimating the reliability of such observational measurements.

The TPOR was developed in conjunction with the Personal Beliefs Inventory and the Teacher Practices Inventory, which attempt to measure teacher beliefs with respect to Dewey's experimentalism.[1] The value of

[1] Bob Burton Brown, The Experimental Mind In Education (New York: Harper and Row, in press with expected publication date September of 1967). See also Brown's "The Relationship of Experimentalism to Classroom Practice." Unpublished Ph.D. thesis. Madison: The University of Wisconsin, 1962.

these three instruments to educational research is not that they measure agreement-disagreement with Dewey's philosophy but that they permit comparable measurements of beliefs and practices in terms of a common theoretical referent. It is this connection with companion measurements of beliefs which differentiates the TPOR from most other instruments for recording observations of classroom behavior. Likewise, its capability for measuring and comparing observed teacher behavior with logically congruent criteria for judging teacher competence gives the TPOR a key function in our "Investigation of Observer-Judge Ratings of Teacher Competence" at the University of Florida, a four-year research project funded by the U. S. Office of Education.

TEACHER PRACTICES OBSERVATION RECORD

The directions for the use of the Teacher Practices Observation Record are as follows:

> The Teacher Practices Observation Record provides a framework for observing and recording the classroom practices of the teacher. Your role as an observer is to watch and listen for signs of the sixty-two teacher practices listed and to record whether or not they were observed, WITHOUT MAKING JUDGMENTS AS TO THE RELATIVE IMPORTANCE OR RELEVANCE OF THOSE PRACTICES.
>
> There are three (3) separate 10-minute observation and marking periods in each 30-minute visit to the teacher's classroom. These are indicated by the column headings I, II, and III. During period I, spend the first 5 minutes observing the behavior of the teacher. In the last 5 minutes go down the list and place a check (✓) mark in Column I beside all practices you saw occur. Leave blank the space beside practices which did not occur or which did not seem to apply to this particular observation. Please consider every practice listed, mark it or leave it blank. A particular item is marked only once in a given column, no matter how many times that practice occurs within the 10-minute observation period. A practice which occurs a dozen times gets one check mark, the same as an item which occurs only once.

> Repeat this process for the second 10-minute period, marking in Column II. Repeat again for the third 10-minute period, marking in Column III. Please add the total number of check marks recorded for each teacher practice and record in the column headed TOT. There may be from 0 to 3 total check marks for each item.

The revised form of the Teacher Practices Observation Record is presented below. It contains 62 items or "signs" of teacher practices. With respect to Dewey's philosophy of experimentalism, 31 of these are positive and 31 are negative. All even-numbered items are positive and all odd-numbered items are negative, making it easy to score the results.

The Teacher Practices Observation Record is usually scored by first totaling the number of check marks for each item, placing either a 0, 1, 2, or 3 in the column headed TOT. Next, the totals for all of the odd-numbered items are reversed, changing 0 to 3, 1 to 2, 2 to 1, and 3 to 0. Then by adding the totals for all items (both the totals for the untouched even or "positive" items and for the adjusted odd or "negative" items) we get a net score. A maximum score of 186 indicates complete experimentalism and a minimum score of 0 indicates complete non-experimentalism. A score of 94 or above indicates the observed teacher practices are more experimental than non-experimental, and a score of 93 or below indicates the opposite.

TEACHER PRACTICES OBSERVATION RECORD

TOT	I	II	III	TEACHER PRACTICES A. NATURE OF THE SITUATION
				1. T makes self center of attention.
				2. T makes p center of attention.
				3. T makes some thing itself center of p's attention.
				4. T makes doing something center of p's attention.
				5. T has p spend time waiting, watching, listening.
				6. T has p participate actively.
				7. T remains aloof or detached from p's activities.
				8. T joins or participates in p's activities.
				9. T discourages or prevents p from expressing self freely.
				10. T encourages p to express self freely.
				B. NATURE OF THE PROBLEM
				11. T organizes learning around Q posed by T.
				12. T organizes learning around p's own problem or Q.
				13. T prevents situation which causes p doubt or perplexity.
				14. T involves p in uncertain or incomplete situation.
				15. T steers p away from "hard" Q or problem.
				16. T leads p to Q or problem which "stumps" him.
				17. T emphasizes gentle or pretty aspects of topic.
				18. T emphasizes distressing or ugly aspects of topic.
				19. T asks Q that p can answer only if he studied the lesson.
				20. T asks Q that is not readily answerable by study of lesson.
				C. DEVELOPMENT OF IDEAS
				21. T accepts only one answer as being correct.
				22. T asks p to suggest additional or alternative answers.
				23. T expects p to come up with answer T has in mind.
				24. T asks p to judge comparative value of answers or suggestions.
				25. T expects p to "know" rather than to guess answer to Q.
				26. T encourages p to guess or hypothesize about the unknown or untested.
				27. T accepts only answers or suggestions closely related to topic.
				28. T entertains even "wild" or far-fetched suggestion of p.
				29. T lets p "get by" with opinionated or stereotyped answer.
				30. T asks p to support answer or opinion with evidence.

TOT	I	II	III	D. USE OF SUBJECT MATTER
				31. T collects and analyzes subject matter for p.
				32. T has p make his own collection and analysis of subject matter.
				33. T provides p with detailed facts and information.
				34. T has p find detailed facts and information on his own.
				35. T relies heavily on textbook as source of information.
				36. T makes a wide range of informative material available.
				37. T accepts and uses inaccurate information.
				38. T helps p discover and correct factual errors and inaccuracies.
				39. T permits formation of misconceptions and overgeneralizations.
				40. T questions misconceptions, faulty logic, unwarranted conclusions.
				E. EVALUATION
				41. T passes judgment on p's behavior or work.
				42. T withholds judgment on p's behavior or work.
				43. T stops p from going ahead with plan which T knows will fail.
				44. T encourages p to put his ideas to a test.
				45. T immediately reinforces p's answer as "right" or "wrong".
				46. T has p decide when Q has been answered satisfactorily.
				47. T asks another p to give answer if one p fails to answer quickly.
				48. T asks p to evaluate his own work.
				49. T provides answer to p who seems confused or puzzled.
				50. T gives p time to sit and think, mull things over.
				F. DIFFERENTIATION
				51. T has all p working at same task at same time.
				52. T has different p working at different tasks.
				53. T holds all p responsible for certain material to be learned.
				54. T has p work independently on what concerns p.
				55. T evaluates work of all p by a set standard.
				56. T evaluates work of different p by different standards.
				G. MOTIVATION, CONTROL
				57. T motivates p with privileges, prizes, grades.
				58. T motivates p with intrinsic value of ideas or activity.
				59. T approaches subject matter in direct, business-like way.
				60. T approaches subject matter in indirect, informal way.
				61. T imposes external disciplinary control on p.
				62. T encourages self-discipline on part of p.

FILM STUDIES

The original 70-item form of the TPOR was used in the spring of 1964 for recording observations of five filmed teaching episodes by a large number of observer-judges at four different teacher education institutions in California, Illinois, New York, and Wisconsin. A year later TPOR observations were repeated on two of these films by the same observer-judges. These data were used to give us information about the consistency-stability reliability of the TPOR.

The teaching episodes observed in this study were originally filmed at Madison, Wisconsin, in the early 1960's. For the purpose of this study 30-minute continuous and uninterrupted segments were cut from unedited films which were 50 to 60 minutes in length. Selection of the films and the segments taken from them was made for purposes of achieving variety in teaching style, and in grade level and subject taught. Teachers in the film were equally well trained (all had master's degrees) and had been selected for filming at the University of Wisconsin as "showcase" teachers. Film #1 was of a ninth-grade French class; Film #2, a seventh-grade mathematics class; Film #3, a fourth-grade unit on "Weather"; Film #4, a ninth-grade speech class; and Film #5, a seventh-grade science class.

The observer-judges were drawn from the faculties of two large midwestern universities and two large state "teachers college-type" schools--one in the east and one in the far west. The observer-judges included student teaching supervisors, education professors, and professors of academic subjects who volunteered their participation in the project. None of them had seen the films or the TPOR prior to the viewing sessions, held separately at the four different campuses over a span of six weeks.

Conditions of the viewing sessions were similar. All observer-judges received the same 10-minute explanation, by the same person, for recording their observations in the TPOR. During the viewing of Film #1 time was called periodically for the observer-judges and lights were switched on and off to make it easier for them to become familiar with the observational procedures and instrumentation. This constituted the sum total of "training" provided the observers. No attempt was made to bring them to any sort of agreement with respect to their recorded observations, nor was any discussion to this effect permitted. Assistance with respect to time and lighting was discontinued after the first film observation, putting the observers "on their own" in every respect.

Table I shows the mean TPOR score given each of the five films by the observer-judges on the first viewing. The French teacher in Film No. 1 was seen as the least experimental and the fourth-grade teacher in Film No. 3 as the most in agreement with Dewey. The range of more than 40 points between the high and low TPOR means indicates the ability of the instrument to differentiate various styles of teaching.

TABLE I

Mean TPOR Scores Given Five Films
by All Observers

Film	No. of Observers	Mean	S. D.
No. 1	130	80.01	13.32
No. 2	124	115.86	16.84
No. 3	119	120.96	22.74
No. 4	119	104.24	17.10
No. 5	67	98.84	12.88

We looked for differences in the TPOR scores given at the four different participating institutions. The location variable was found to have little or no influence. Using Scheffe's comparisons, no statistically significant differences were found among the TPOR means given at the various locations for Film Nos. 1, 2, 4, and 5. The only statistically significant differences were found between California and each of the other three locations on Film No. 3.

We also looked for differences in the TPOR scores given by the three major occupational classifications of observer-judges--college supervisors of student teaching, education professors, and academic professors. No statistically significant differences were found between any of these groups for Film Nos. 1, 2, 4, and 5. The only statistically significant differences were found between supervisors of student teaching and both education and academic professors on Film No. 3.

TPOR means were also examined in relation to the evaluative judgments made about the quality of teaching observed in the films. Table II shows an interesting pattern of correlation between TPOR scores and ratings given each film. While this could mean that the TPOR scores were influenced by how much the observer liked what he saw, the converse is more likely true. The wide differences in TPOR means within each of the evaluative categories are evidence that the correlation between TPOR scores and ratings is relative within the limits describing each individual film. In this study, a given TPOR score did not guarantee a "good" or "bad" rating, even though in every case the higher the rating, the higher the TPOR mean score.

TABLE II

The Relationship Between TPOR Means and Evaluative Ratings of Five Filmed Teaching Episodes

	Evaluative Ratings					
	A	B	C	D	E	F
Film	Out-standing	Very Good	Good	Fair	Poor	In-competent
No. 1	88.64 (11)	82.21 (48)	79.45 (33)	67.89 (9)	--- (0)	--- (0)
No. 2	126.47 (19)	118.57 (56)	109.19 (21)	--- (0)	--- (0)	--- (0)
No. 3	138.19 (27)	119.32 (38)	109.91 (23)	85.00 (1)	--- (0)	--- (0)
No. 4	110.69 (29)	106.86 (43)	93.73 (11)	76.50 (2)	75.50 (2)	--- (0)
No. 5	115.56 (9)	103.39 (13)	96.00 (23)	88.67 (6)	65.00 (1)	--- (0)

Statistically significant differences beyond the .05 level (using Scheffe's comparison procedures) were found for the following pairs of means:

Evaluative Category A: Films 1 and 2, 1 and 3, 3 and 4, (1 and 4, 1 and 5 were very close)

Evaluative Category B: Films 1 and 2, 1 and 3, 1 and 4, 1 and 5

Evaluative Category C: Films 1 and 2, 1 and 3 (1 and 5 were close)

Film No. 3: Category A and B, A and C

Mass observations of films are expensive and administratively difficult to arrange. For these reasons repeated observations the second year could be obtained on only two of the five films. Film No. 1 was eliminated because it had been used as the "training" film and the conditions of the first viewing could not be simulated. Film No. 3 yielded a wide discrepancy between the scores given it in California and those given it at the other three locations, which we thought might be due to the artificial conditions under which it was filmed. Film No. 5 had not been observed at all four institutions. This left No. 2 and No. 4, which were selected by elimination for the second viewing. It was possible to obtain repeated TPOR scores on these two films by only a portion of those who observed the first viewings.

Table III shows a fairly substantial difference between TPOR means recorded for the first and second viewings of Film No. 2. While this difference raises some questions about stability, both means for this subgroup of 69 observers lie well within one standard deviation of the mean of 115.86 for 119 first-viewing observers which may simply demonstrate the normal variability of TPOR scores. The differences between TPOR scores for the first and second viewings of Film No. 4 are very small.

TABLE III

Mean TPOR Scores Given Films On
Repeated Observations One Year Apart

Film	Viewing	No. Observers	Mean	S. D.
No. 2	1st	69	122.22	20.52
No. 2	2nd	69	109.81	18.31
No. 4	1st	72	107.15	17.15
No. 4	2nd	72	105.14	18.12

RELIABILITY

In order for anyone to place confidence in the scores obtained with the TPOR, its reliability as a measuring instrument must be established. There are three major problems involved in doing this:

1. Selecting types (or definitions) of reliability appropriate to the instrument and the purposes for which it is designed.
2. Selecting a meaningful measure (or yardstick) of reliability once the type is specified.
3. Selecting a good estimator of a given measure to give an estimate of reliability based on experimental data.

Reliability can be a tricky concept. We know that reliability always refers to consistency throughout a series of measurements, and that it is usually expressed in terms of something called reliability coefficients. Rarely do we make clear what kind of consistency has been figured. Although everybody in educational research reads reliability coefficients, few seem to really understand (or care) what these mean or how they were obtained. All that matters is that they be high. Once the standard for "highness" has been debated and denoted, then surpassed or fallen short of, what more is there to say about reliability?

There are many different kinds of reliability to be considered. Thorndike speaks of approaching the study of reliability from two quite different viewpoints. One approach is to be concerned about the actual or absolute magnitude of errors of measurements. In this case reliability is expressed in terms of the variability of scores obtained by repeated testing of the same individual, and is based on a statistic called standard error of measurement. Another approach can be made in terms of the consistency with which individuals maintain the same relative position in the total group on repetition of a measurement procedure. In this case consistency is expressed in terms of the correlation between two sets of scores, called the _coefficient of reliability_.[2] As a further example, Cronbach points out that not all reliability coefficients reveal the same or even comparable information. He refers to "comparable-forms," "split-half," and "test-retest" reliability coefficients as ways to get at different aspects of reliability. The first is a "coefficient of equivalence and stability," the second a "coefficient of equivalence" only and the third a "coefficient of stability."[3] Furthermore, we have something called _internal consistency_ or _item reliability_ which assesses test homogeneity, or the extent to which all items measure the same attribute. This, of course, is a horse of still another color. All of which makes the use of the term "reliability" meaningless without some sort of further differentiation and definition.

Dealing adequately with already difficult concepts of reliability becomes even more complex when one turns from consideration of tests of achievement and intelligence, and the like, to the measurement of classroom behavior by systematic observation. The question of the reliability of the observers and the recording of their observations must be added to the problem. In the past most observational studies have limited their

[2] Robert L. Thorndike, "Reliability," Chapter 15 in _Educational Measurement_, E. F. Lindquist, Editor. (Washington, D. C.: American Council on Education, 1950), pp. 560-61.

[3] Lee J. Cronbach, _Essentials of Psychological Testing_, Second Edition, (New York: Harper & Brothers, 1960), pp. 136-142.

study of reliability to computing the correlation between two sets of observations or to figuring the percent of agreement between observers.

Keeping this tradition, in part, we computed the correlations between the TPOR scores obtained from the repeated observations of Films No. 2 and No. 4. It is curious to note in Table IV that the correlations of the columns (10-minute observation periods) within each film observation are very high, but the correlations between the 1964 and 1965 observations are very low. The first indicates that the observers tended to maintain

TABLE IV

Correlation of TPOR Scores Obtained from
Repeated Observations of Films

FILM NO. 2

TPOR Column		1964 Observation				1965 Observation			
		TPOR Column				TPOR Column			
		1	2	3	TOT	1	2	3	TOT
1964 Obser-vation	1	1.00	.79	.69	.89	.36	.25	.12	.27
	2	--	1.00	.81	.95	--	.29	.16	.31
	3	--	--	1.00	.92	--	--	.20	.29
	TOT	--	--	--	1.00	--	--	--	.32
1965 Obser-vation	1	--	--	--	--	1.00	.61	.55	.80
	2	--	--	--	--	--	1.00	.81	.93
	3	--	--	--	--	--	--	1.00	.90
	TOT	--	--	--	--	--	--	--	1.00

FILM NO. 4

TPOR Column		1964 Observation				1965 Observation			
		TPOR Column				TPOR Column			
		1	2	3	TOT	1	2	3	TOT
1964 Obser-vation	1	1.00	.75	.52	.86	.32	.36	.25	.34
	2	--	1.00	.71	.93	--	.46	.52	.52
	3	--	--	1.00	.85	--	--	.67	.57
	TOT	--	--	--	1.00	--	--	--	--
1965 Obser-vation	1	--	--	--	--	1.00	.79	.71	.90
	2	--	--	--	--	--	1.00	.83	.95
	3	--	--	--	--	--	--	1.00	.92
	TOT	--	--	--	--	--	--	--	1.00

the same relative position in the group throughout the viewing of a single film on a given day. The second indicates that sizeable shifts in these positions took place during the intervening year. In other words we got good consistency within one occasion or viewing, and again within another, but poor stability between two widely separated occasions. One must keep in mind, however, that such reliability coefficients normally decline proportionately with the length of time between "tests." Had the repeat observations been made only a month or so apart we might expect considerably higher correlations.

Even so, correlation of two sets of scores by a number of different observers is not likely to be a very accurate estimate of reliability. It is difficult to make arrangements for large numbers of observers to view the same classroom on two different occasions, or to control variations between those occasions. Likewise, the number of classrooms observed on two different occasions by two different observers is likely to be small. In either case, the size of the N determines the precision of the correlation coefficient, and since the N of even well-financed observational studies rarely exceeds 100 the confidence intervals for the coefficients are extremely wide. Furthermore, such correlations are usually based on total scores which ignore variations in scoring individual items or categories. It is possible to obtain a perfect correlation of total scores when the reliability for the items is zero. If on a 70-item "sign" system, for example, the 35 odd-numbered items are marked "+" and the 35 even-numbered items are marked "0" on the first observation, and then exactly reversed on the second observation, identical total scores will be obtained and used to produce a deceivingly perfect reliability correlation.

Percent of agreement between observers tells almost nothing about the accuracy of the scores obtained. It is entirely possible to find observers agreeing 99 percent in recording behaviors on an instrument whose item or category consistency is very poor. Reliability can be low even though observer agreement is high for several reasons. For example, observers might be able to agree perfectly that a particular teaching practice occurred in a classroom, yet if that same practice occurs equally, or nearly so, in all classrooms, the reliability of that item as a measure of differences between teachers will be zero. Near-perfect agreement could also be reached about the percentage of time a number of teachers employed certain categories of behavior; but if every teacher sharply reversed these percentages from period to period or day to day, the reliability of these categories would be zero. Errors arising from variations in behavior from one situation or occasion to another can far outweigh errors arising from failure of two observers to agree exactly in their records of the same behavior.

Yet, the reliability of most instruments for systematically recording the behavior of teachers, including Flanders' well-known Classroom Interaction Analysis, requires a high percent of observer agreement. "Between-observer" agreement has become almost a cardinal principle in planning observational studies. According to Medley and Mitzel a sample of classrooms from the population to be studied should be visited by trained recorders using the observational instrument in the same way it will be used in any subsequent study. In order to study the "objectivity" of the items, i.e., how closely observers agree in recording identical behaviors, at least two recorders should be present on each visit, sitting in different parts of the room and making independent records. In order to be able to estimate how stable the two records based on different visits will agree, each class should be visited at least twice. To recapitulate, in their words, "*c* teachers are visited in *s* situations by a team of *r* recorders. In studying the reliability of a scale with *i* items on it, the total number of scores to be analyzed will be *cris*."[4]

[4] Donald M. Medley and Harold E. Mitzel "Measuring Classroom Behavior by Systematic Observation," Chapter 6 in Handbook of Research on Teaching, N. L. Gage, Editor (Chicago: Rand McNally & Company, 1963), p. 309.

To match this rigorous plan for data collection Medley and Mitzel have taken the classic definition of reliability, $\rho_{xx} = \frac{\sigma_\tau^2}{\sigma_x^2}$ and applied it to measurements of classroom behavior. In this definition, true variation, σ_τ^2 is defined to be the variation of the total score for any class (teacher) when the effects of recorders (observers), items on the scoring instrument, and situations (viewings or visits) have been removed. The true variation plus "error," σ_x^2, is defined to be the variation of the total scores for any class, including variation contributed by items on the scoring instrument, recorders, situations and random error. The smaller the effect of the recorders, items, and situations for a class total, the higher the reliability coefficient will be. In other words, if the instrument has high reliability the scoring of the class or teacher is relatively free of the effects of recorders, items, or the different situations under which the scoring was done, and as such, reflects a "good" or reliable instrument.[5]

[5] Ibid.

In seeking a design for estimating the reliability of TPOR observations, we closely examined the four-way analysis of variance model suggested by Medley and Mitzel. While we found it to be a sound approach to reliability estimation, it may not be entirely appropriate for analyzing the data obtained in the film study described above. For instance, in the simple example given by Medley and Mitzel in the Handbook of Research on Teaching, page 316, where one item is used to score 24 classes (teachers) observed during four situations by two recorders (observers), the reliability coefficient is estimated by:

$$\rho_{xx} = 1 - \frac{MS_{cxr}}{MS_c}$$

Where MS_{cxr} is the mean square for classes x recorders obtained from the analysis of variance table and MS_c is the mean square for classes obtained from the analysis of variance table. The coefficient of reliability in this case actually reflects not instrument reliability, but rather, recorder or observer reliability. When MS_{cxr} is large, it indicates an inconsistency on the part of the observers to score the classes in the same way, which in turn causes ρ_{xx} to be small. In like manner, a very small value of MS_{cxr} reflects consistency in scoring, in which case ρ_{xx} will be large.

Training of the observers undoubtedly would bring them into agreement with respect to recording or scoring identical behaviors, which would be reflected in a higher reliability coefficient, ρ_{xx}. However, in the previously described film study in which the TPOR was tried out, no attempt was made to train the observers. To the contrary, we deliberately tried to preserve the differences among observers by selecting them from varying occupational groups, from varying sizes of institutions with varying orientations to teacher education, and from varying parts of the country. We wanted to test the reliability of the TPOR under uncontrolled field conditions to see what value it might have in the hands of the differing kinds of people who carry out the everyday responsibilities for teacher education in America. Hence, the component of variance due to the observers' variability in our study would cause σ_x^2 to be large compared to σ_τ^2, resulting in a small $\rho_{xx'}$. We did not get as much observer variability as might have been expected, however. When the Medley-Mitzel model was adapted to fit our film study data the TPOR observations were found to have a modest but substantial reliability coefficient of .57.

In the analysis of variance example cited above it should also be noted that two of the variables of interest, viz., classes and situations, had but one degree of freedom each. This being the case, "poor" estimates of the components of variance could result. In fact, the components of variance could be estimated to be zero (which happens in many cases). Also, since the estimate of ρ_{xx} would consist of the ratio of linear combinations of mean squares, the bounds of error on this estimate could be exceedingly large.

The unsuitability of the Medley-Mitzel model for our data results primarily, however, from the fact that it stresses "between-observer" variability rather than "within-observer" variability. This is a philosophical rather than a statistical issue. Reliability coefficients which reward high agreement between observers implies that we should seek a single, uniform, "objective" system for observing and classifying teaching behavior. From the point of view of the framework underlying the development of the TPOR, objectivity in perceiving and quantifying such behavior is neither possible nor desirable. "Between-observer" agreement may not only encourage a false sense of confidence with respect to the accuracy of measurements, but also gives us a false sense of "objectivity" regarding the observations. A team of observers can be brainwashed to the point of near-perfect agreement, but this does not erase the possibility that instead of several differing "subjective" judgments, they now make only one. Therefore, we sought another mathematical definition of reliability, one which is concerned primarily with "within-observer" variability.

We reasoned that if having scored a given filmed teaching situation, the same observer-judge were to score the same teaching situation again in the same way, then we could say the observer-judge's scoring was reliable. Hence, a definition for "within-observer" reliability for a given observer-judge and film was devised as follows:

	Viewing		
Items	1	2	$d_i = x_{1i}-x_{2i}$
1	x_{11}	x_{21}	d_1
2	x_{22}	x_{22}	d_2
3	x_{13}	x_{23}	d_3
⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	d_n

Consider the variances of the differences d_i, where

$$d_i = x_{1i}-x_{2i}$$

If the scores are independent, i.e., the judge is not consistent, or in fact marks by chance, then

$$\begin{aligned} V(d_i) &= V(x_{1i}-x_{2i}) \\ &= V(x_{1i}) + V(x_{2i}) \\ &= \sigma^2 + \sigma^2 \\ &= 2\sigma^2 \quad \text{(or 2 Var(x))} \end{aligned}$$

However, if the judge is consistent from viewing to viewing, his 2 scores should be positively correlated and now

$$\begin{aligned} V(d_i) &= V(x_{1i}- x_{2i}) \\ &= V(x_{1i}) + V(x_{2i}) - 2 \operatorname{Cov}(x_{1i},x_{2i}) \\ &= 2\sigma^2- 2\sigma_{12} \end{aligned}$$

or

$$V(d_i) = \sigma_d^2 = 2\sigma^2- 2\sigma_{12}$$

It is noted that the following assumptions are made in the above discussion:

1) The variance of each item score is the same for all items over viewings; i.e.,

$$V(x_{ij}) = \sigma^2 \quad \text{for} \quad \begin{array}{l} i=1,2 \\ j=1...n \end{array}$$

2) Under the complete randomness assumed under chance scoring, each value of x is assumed to have equal chance of being selected; hence

$$p(x)= \frac{1}{k}$$

where k is the number of choices available.

Now we define for judge j and film f,

$$\rho_{jf} = 1 - \frac{\sigma_d^2}{2\sigma^2}$$

where $\sigma_d^2 = \mathrm{Var}(d_i)$ $i=1\ldots n$

$\sigma^2 = \mathrm{Var}(x_{ij})$ $i=1,2$

$j=1\ldots n$

However, under the assumptions of a random choice by tbe judge, σ^2 becomes a constant, computed as

$$\sigma^2 = \sum_x (x-\mu)^2 p(x)$$

We calculate the sample value of s_d^2 and use it to estimate σ_d^2. Hence we are working with a statistic

$$r_{jf} = 1 - \frac{s_d^2}{2\sigma^2}$$

Now, if there is in fact high positive correlation of the scoring from viewing 1 to viewing 2, then

s_d^2 will be small (i.e., s_{12} will be large)

and

r_{jf} will be close to 1.

If the scoring from viewing to viewing is in fact independent and really associated with a chance event, then

s_d^2 will be of the magnitude of $2\sigma^2$ (i.e., s_{12} will be small; close to zero)

and

r_{jf} will be close to 0.

The coefficient r_{jf} will theoretically be in the interval (0,1) where a maximum value of one implies absolute correlation, while a minimum value of zero implies the same scoring could have happened by chance, hence no reliability. However, the possibility of $r_{jf}<0$ exists because there is a non-zero probability that the scorings will be negatively correlated and this may cause s_d^2 to be greater than σ^2; this in turn causing $r_{jf} < 0$.

Worth mentioning is the fact that this statistic uses a larger than expected variance σ^2, as a yardstick against which the judge's variation from viewing to viewing is compared. This is because one would expect a judge to select the extremes in scoring an item less frequently than scores near the center of the scale; such scoring would likely yield a variance smaller than that implied by a completely random selection. This yardstick could, in effect, cause the coefficient r_{jf} to be depressed as compared with other measures of reliability.

Using the above formulation the "within-observer" reliability of TPOR scores was computed for the two filmed teaching situations on which repeated viewings were made a year apart. Table V shows eight reliability coefficients ranging between .48 and .62.

TABLE V

"Within-Observer" Reliability Coefficients for TPOR Scores on Repeated Viewings of Films

FILM NO. 2

N = 69

TPOR Column	r_{jf}	error
TOT	.48	.0255
1	.57	.0177
2	.51	.0194
3	.51	.0177

FILM NO. 4

N = 72

TPOR Column	r_{jf}	error
TOT	.52	.0191
1	.56	.0182
2	.57	.0244
3	.62	.0171

These coefficients of reliability, as is the case with those obtained using the Medley-Mitzel model, reflect observer reliability rather than instrument reliability. Observer reliability is always subject to variations in the selection and training of people and the control of conditions under which they use an instrument. People and conditions can be "improved" in subsequent studies, but once they are "out," instruments rarely are. So it is important to know about the internal consistency of the instrument, its item reliability--which tells us something of its potential in the hands of reliable observers.

Table VI shows the results of submitting the film study data to the Kuder-Richardson formulation for measuring item reliability. If each item is highly correlated with every other item on the instrument, then the instrument has good item reliability or internal consistency. The fact that the TPOR scores yielded uniformly high internal reliability coefficients is not surprising in light of the fact that throughout their development the TPOR, TPI, and PBI underwent repeated RAVE analysis, an iterative procedure which yields a set of item response weights which maximize the internal consistency of inventories.[6]

[6] Ronald Ragsdale and Frank B. Baker, The Method of Reciprocal Averages for Scaling of Inventories and Questionnaires: A Computer Program for The CDC 1604 Computer, (Mimeographed, Laboratory of Experimental Design, Department of Educational Psychology, University of Wisconsin, Madison).

TABLE VI

TPOR Internal Consistency Reliability Coefficients

			TPOR Columns			
Film	Viewing	N	1	2	3	TOT
No. 1	1st	158	---	---	---	.86
No. 2	1st	69	.79	.81	.83	.93
"	2nd	69	.77	.81	.79	.91
No. 3	1st	140	---	---	---	.93
No. 4	1st	72	.76	.77	.78	.90
"	2nd	72	.76	.78	.77	.91
No. 5	1st	84	---	---	---	.85

In summary, we wish to emphasize that the TPOR was developed for wide-scale field use by "untrained" observers in the study of teaching behavior in relation to philosophic and education beliefs. Instead of trying to "train out" the pluralistic biases in the perceptions of our observer-judges, we deliberately left them alone, took them just as they came, and tried to include them and take them into account as we analyzed the results obtained. This analysis, of course, awaits reporting elsewhere. In this paper we are concerned only with reporting, in the context of a discussion of problems involved in defining and measures of reliability, the reliability data obtained from experimental use of the TPOR. Having submitted this instrument to the hazards of uncontrolled use by uncontrolled observers, and then submitting it to the severest statistical procedures we could find, it came out with the following score card: (1) Correlation of observers' total scores within a given film viewing--VERY GOOD, (2) Correlation of observers' total scores between repeat film viewings one year apart--POOR to FAIR, (3) Between-observer reliability--FAIR, (4) Within-observer reliability--FAIR, (5) Internal consistency reliability--VERY GOOD.