DOCUMENT RESUME

SP 002 275

ED 028 123

By-Webb, Jeaninne Nelson; Brown, Bob Burton
Establishing Reliability and Validity Estimates for Systematic Classroom Observation.
Pub Date [69]
Note-11p.; Presented at the 1969 American Educational Research Association meeting, February 1969, Los Angeles, California.
EDRS Price MF-$0.25 HC-$0.65
Descriptors-Beliefs, *Classroom Observation Techniques, Educational Experiments, Educational Researchers, Evaluation, *Reliability, Research Skills, Training, *Validity
Identifiers-PBI, Personal Beliefs Inventory, Teacher Practices Inventory, Teacher Practices Observation Record, TPI, TPOR

A study was designed to (1) compare two types of reliability in the observation of teachers' behavior, (2) explore the relationship between observer reliability and the validity of their systematic classroom observations, and (3) investigate the effects of training, observer beliefs, and the passage of time on reliability and validity estimates. Subjects were 32 experienced elementary school teachers, 16 of whom attended five 2-hour training sessions in the use of the Teacher Practices Observation Record (TPOR) and all of whom took the Personal Beliefs Inventory and the Teacher Practices Inventory to measure beliefs. Both groups viewed two films of classroom teacher behavior and used the TPOR to record observed behavior twice, once 10 weeks after training and again 10 weeks after that. Reliability coefficients were computed for between-observer agreement and within-observer agreement (the stability of an observer's response over time). These and criterion validity coefficients were used as responses in linear multiple regression analysis. Conclusions were that if observers share a common perceptual framework (as these did), between-observer agreement can be achieved easily with little or no training, but within-observer reliability is difficult to achieve. Therefore, training should focus on establishing the reliability of the individual observer. (JS)

ESTABLISHING RELIABILITY AND VALIDITY ESTIMATES FOR SYSTEMATIC CLASSROOM OBSERVATION

by

Jeaninne Nelson Webb
College of Education
University of Alabama

and

Bob Burton Brown
College of Education
University of Florida

Introduction

All systems of classroom observation share one common element: their dependence upon the observers who use them. Too often, little attention has been given to procedures used in training observers and the methods by which data is computed and reported in regard to observer reliability and validity. Yet without an estimation of the accuracy and relevancy of observation data collected, little confidence can be placed in the findings they produce.

Problem

The purposes of the study were (1) to compare two types of reliability in the observation of teachers' behavior, (2) to explore the relationship between observer reliability and the validity of their systematic classroom observations and (3) to investigate the effects of training, measured observer beliefs, and the passage of time on reliability and validity estimates.

Procedures

Instrumentation. Scores obtained by the employment of an observation system, the Teacher Practices Observation Record (TPOR),

were used to establish reliability and validity estimates. The TPOR is a 62-item sign system which measures the instructional practices of a teacher in terms of agreement-disagreement with John Dewey's experimentalism. The observation is recorded during a 30-minute session which is divided into three ten-minute observation and marking periods; each of the sixty-two items is to be considered and then checked if the described behavior occurs during the period. Thus the observer is required to make 186 discriminations as to the presence or absence of the described practices during the total observation period. From the observation a descriptive record of teaching behavior can be reported in the form of a numerical score ranging from 0 to 186. A TPOR score of 93 or above indicates teacher behavior in greater agreement than disagreement with experimentalism, below that to be in greater disagreement than agreement.

Through recognition that observer biases and subjectivity will color records of classroom behavior, two instruments were used to measure the beliefs of subjects used as observers. The Personal Beliefs Inventory (PBI) and Teacher Practices Inventory (TPI)[1] were developed to be used in conjunction with the Teacher Practices Observation Record and measure fundamental philosophic and educational beliefs. High scores indicate agreement with experimentalism; low scores indicate rejection of Dewey's philosophy.

---

[1]For an account of the development of these instruments and the Teacher Practices Observation Record, see Bob Burton Brown. The Experimental Mind in Education. New York: Harper and Row, 1968.

Subjects. The subjects of the study were thirty-two experienced female elementary teachers selected from a rural Florida county. Sixteen of the subjects were trained in the use of the observation system; they comprised the trained group. The sixteen remaining subjects received no training.

Training Procedures. The training of observers consisted of five two-hour training sessions held over a four-week period; films of teachers in unrehearsed classroom situations were used for training purposes. Provision was made to give observers immediate feedback on agreement. Efforts were made by the trainer to encourage the observer subjects (1) to achieve agreement in their responses to the observation instrument and (2) to record behavior in terms of the theoretical basis of the instrument.

Data Collection. Two films of classroom teacher behavior (A and B) were used for data collection purposes. Each group of subjects viewed the two films and recorded the observed behavior twice, once approximately ten weeks after training had been completed and then again ten weeks after the first viewing session. The TPOR scores obtained in the two viewing sessions were used in the analysis of the data.

Data Analysis. The data were first used to compute two types of reliability coefficients: (1) Between-observer, the agreement between observers of the same teacher behavior and computed as a percent of agreement. This coefficient is a ratio of the number of responses to which observers agree to the total number of responses possible, and (2) Within-observer, the stability of an individual observer's responses to the same behavior over a period of time.

These coefficients were computed by techniques developed by Brown, Mendenhall, and Beaver.[2] In addition, criterion validity coefficients were developed by comparing observers' scores with criterion scores. Criterion scores were composite scores given the films by the trainer and the author of the Teacher Practices Observation Record. The validity coefficient was computed using the same procedures as the within-observer reliability coefficient.

The coefficients established by these procedures were used as responses in linear multiple regression analysis to investigate the effects of training, the effects of measured beliefs and the effects of time on the reliability and validity of observers' observation scores. Lastly, the relationships between the validity coefficients and reliability of observations were examined.

Findings

Between-Observer Reliability Coefficients. Between-observer reliability coefficients were computed for each film for each viewing and for the variables under investigation, the effects of training and measured beliefs, and are reported in Table 1. These coefficients ranging from .77 to .86 are comparable with those reported for other observation instruments and are remarkably uniform. The single identifiable general trend was that untrained observers achieved slightly higher coefficients than trained observers.

---

[2] Bob Burton Brown, William Mendenhall, and Robert Beaver, "The Reliability of Observations of Teachers' Classroom Behavior," Journal of Experimental Education, 36:1-10, Spring, 1968.

TABLE 1

BETWEEN-OBSERVER RELIABILITY COEFFICIENTS

	Trained - Observers			Untrained - Observers		
	Low* Belief Scores	High* Belief Scores	Total	Low* Belief Scores	High* Belief Scores	Total
	N=8	N=8	N=16	N=8	N=8	N=16
Film A Viewing 1	.80	.81	.80	.83	80	.80
Film A Viewing 2	.78	.79	.77	.82	.85	.82
Film B Viewing 1	.82	.82	81	.83	85	.83
Film B Viewing 2	.81	.81	.79	.84	86	.84

*This is a relative classification within the groups; all subjects belief scores fell within a fairly narrow range.

Also the trained observers' agreement tended to decrease slightly over time. The variable of beliefs seemed to have no effect, and training had precious little effect, on agreement between observers.

Clearly, for the Teacher Practices Observation Record, neither training nor beliefs have a great effect on between-observer agreement. It could be that training was ineffectual in contributing to between-observer agreement. Efforts were made in training sessions to encourage observers to record behavior in terms of the theoretical basis of the instrument, even at the expense of increasing agreement.

This might have counterbalanced any tendency for the trained group to reach higher agreement than the untrained group. Another factor which could well have affected these results was the subjects themselves. The members of both groups were remarkedly similar to one another in beliefs, sex, occupation and, in general, shared the same socio-economic background. The untrained group of observers shared so many attitudes and common experiences they tended to perceive behavior in much the same frame of reference. Thus, for them, a consensus in the perception of teaching behavior had been achieved by environmental factors long before their viewing of the films. A third factor which may account for the similarity of agreement in both trained and untrained groups is the observation instrument itself. The Teacher Practices Observation Record was designed to be used by untrained observers; in its development, only items which described behavior in clear and concise terms were included in the final form. Thus, the composition of the instrument itself leads to agreement of responses of observers who share similar perceptual frameworks.

Within-Observer Reliability Coefficients. Through the comparison of responses of each observer to each film for the first and second viewings, individual within-observer coefficients, the stability of an observers scores, were developed and can be found in Table 2. The coefficients range from .34 to .77 for the trained group and from .41 to .90 for the untrained group. Mean within-observer coefficients for trained and untrained groups were very uniform and are shown in Table 3. No variables could be identified which would even partially

TABLE 2

WITHIN-OBSERVER RELIABILITY COEFFICIENTS

Trained			Untrained		
Observer ID No.	Film A	Film B	Observer ID No.	Film A	Film B
1	.65	.73	37	.58	.63
2	.64	.77	38	.58	.81
3	.42	.60	39	.51	.45
8	.37	.34	40	.52	.71
9	.73	.70	41	.51	.53
10	.57	.72	42	.57	.57
13	.63	.61	43	.65	.52
15	.60	.73	44	.69	.81
19	.68	.48	45	.57	.53
21	.46	.56	46	.54	.55
24	.59	.67	47	.41	.48
25	.54	.62	48	.78	.90
28	.63	.72	49	.67	.53
29	.59	.51	50	.47	.64
31	.52	.57	51	.60	.62
34	.47	.56	52	.71	.75

account for the wide variance between individual coefficients. There is no question that observers do vary greatly in the stability to which they respond to teaching behavior over time; however the variables of training and beliefs did not seem to influence this stability for the subjects under investigation. The only factor which did seem to affect the within-observer reliability coefficient was the film itself. Observers responded in a more stable manner to Film B than to Film A.

Criterion Validity Coefficients. Individual validity coefficients for each film for each session were computed and are reported in

TABLE 3

MEAN WITHIN-OBSERVER RELIABILITY COEFFICIENTS

	Trained Observers			Untrained Observers		
	Low Belief Scores	High Belief Scores	Total	Low Belief Scores	High Belief Scores	Total
	N=8	N=8	N=16	N=8	N=8	N=16
Film A	.54	.60	.57	.59	60	.59
Film B	.61	.62	.62	.60	66	.63

Table 4. Mean coefficients appear in Table 5. The criterion validity coefficients for these subjects were low with wide variability. Within these general limitations variables were identified which would account for a statistically significant amount of the variance between coefficients. The multiple regression analysis indicated that the interaction of training and belief variables affected the validity of subjects observations. The effects of training on observers more in agreement with experimentalism had a tendency to produce higher validity coefficients. This effect decreased over time. The training of observers less in agreement with experimentalism had a slightly negative effect on validity. This effect increased over time

Comparison of Coefficients. A comparison was made of the relationship between the validity coefficients and the within-observer reliability coefficients. A significant relationship was identified;

TABLE 4

CRITERION-OBSERVER VALIDITY COEFFICIENTS

Trained					Untrained				
	Film A		Film B			Film A		Film B	
Observer ID No.	Viewing 1	Viewing 2	Viewing 1	Viewing 2	Observer ID No.	Viewing 1	Viewing 2	Viewing 1	Viewing 2
1	.54	.42	.36	.39	37	.35	.50	.36	.24
2	.54	.46	.61	.66	38	.28	.32	.42	.42
3	.32	.21	.28	.32	39	.41	.20	.22	.26
8	.29	.29	.24	.32	40	.38	.55	.40	.55
9	.68	.54	.25	.43	41	.31	.32	.10	.40
10	.35	.46	.39	.46	42	.41	.52	.43	.47
13	.32	.24	.41	.23	43	.57	.59	.45	.31
15	.53	.42	.49	.39	44	.32	.40	.32	.38
19	.35	.34	.29	.17	45	.44	.34	.44	.33
21	.45	.29	.43	.34	46	.30	.36	.41	.34
24	.48	.36	.41	.49	47	.19	.38	.27	.34
25	.25	.31	.41	.42	48	.35	.42	.46	.57
28	.59	.56	.54	.49	49	.38	.40	.32	.42
29	.36	.47	.37	.41	50	.45	.24	.26	.20
31	.41	.25	.31	.21	51	.35	.38	.25	.36
34	.42	.32	.23	.22	52	.35	.28	.29	.31

observers who are more consistent in their recording of observations of the same behavior over a period of time also tend to make more valid observations. This relationship is slightly accentuated if the observer has been trained.

Conclusions

For those who gather classroom behavioral data, through systematic observation, between-observer reliability needs close examination.

TABLE 5

MEAN CRITERION - OBSERVER VALIDITY COEFFICIENTS

	Trained Observers			Untrained Observers		
	Low Belief Scores	High Belief Scores	Total	Low Belief Scores	High Belief Scores	Total
	N=8	N=8	N=16	N=8	N=8	N=16
Film A Viewing 1	.40	.46	.43	41	[illegible]	39
Film A Viewing 2	.34	.42	.38	.40	[illegible]	39
Film B Viewing 1	.39	.36	.38	.33	[illegible]	34
Film B Viewing 2	.40	.38	.39	[illegible]6	.38	.37

The classic method of obtaining dependable evidence of what happens in a classroom has been to train observers to use some type of observation instrument, rating scale or check list. The prime purpose of training has been to achieve agreement between observers as to the behaviors they are recording. Thus when the time arrives that observers can agree on what to label the behavior under question, they are considered trained and dependable to gather accurate and relevant information. The data suggest that this is far from the case. If one can find observers who share common perceptual frameworks, agreement can be achieved easily, with little or no training. It is

possible by selecting a fairly uniform sample to find observers who can easily agree, but to get them to observe behavior within a particular theoretical framework is a far more difficult task

Within-observer reliability would seem a far more useful concept for both practical and theoretical reasons. Observations of classroom behavior are expensive and time consuming methods of procuring data. It is difficult and prohibitive in cost to send more than one observer into a classroom to collect data. Therefore, that an observer remain consistent in recording behavior as he moves from classroom to classroom is of more importance than that he reaches agreement with other observers at some point in time, providing he records data in a manner relevant to the instrument he is using.

This leads to the problem of the validity of the observations he makes. It would seem from the literature devoted to systematic classroom observation that validity has been assumed to be achieved automatically with between-observer agreement. If problems of observer validity have been entertained, they have not been reported. No one seems to have squarely faced the factors involved in the validity of classroom observations--do they really measure what they propose to measure?

The study has made an approach to answering the question by attempting to establish criterion validity of observations of teaching behavior. The training procedures used were regrettably less then effective, yet it seems a step in the right direction. The primary purpose of training should be to establish the validity of the observer; the data indicates that, with validity, within-observer reliability will follow.