DOCUMENT RESUME

ED 236 545 CS 007 306

AUTHOR Willson, Victor L.; And Others
TITLE Sources of Variation That Affect the Reliability of Reading Classroom Observation Measures. Instructional Research Laboratory Technical Series #R83004.
INSTITUTION Texas A and M Univ., College Station. Instructional Research Lab.
PUB DATE [82]
NOTE 19p.
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Communication; *Classroom Observation Techniques; *Data Collection; *Error of Measurement; Reading Instruction; *Reading Research; *Reliability; Research Design; *Research Methodology; Research Problems; Teacher Effectiveness

ABSTRACT

Given the complexity of classroom reading observations, maintenance of measurement reliability is a concern to researchers. Sources of variation contributing to the unreliability of many measurements may be either (1) lasting and specific, (2) lasting and general, (3) temporary and specific, or (4) temporary and general. Lasting-specific sources of variations, such as observers' attitudes toward the coding system and differences between training conditions and actual classroom settings, can be controlled through appropriate criteria for observer selection and training procedures that are content specific to the actual observation tasks. Lasting-general sources of variation include observers' observation skills and the supervision condition, which can be controlled, in part, by researchers and observers for discussing recently completed observations. Although temporary-specific sources of variation, such as observer fatigue, attention span, and memory, are not always under experimenter control, they can be managed with some success by such techniques as regulation of observation periods. Temporary-general sources, including location, surroundings, and physical condition of the observer, are more amenable to researcher control. If an observer is ill, for example, the observation session can be cancelled. The systematic consideration of these sources of variation is critical to understanding classroom interaction and behavior. (MM)

Sources of Variation that Affect the Reliability of Reading Classroom Observation Measures

Victor L. Willson, Associate Professor
Department of Educational Psychology
Texas A&M University
College Station, TX 77843
409/845-1808

Nancy G. Mangano, Assistant Professor
Kansas State University

William H. Rupley, Associate Professor
Texas A&M University

IRL Technical Paper Series
#R83004

Victor L. Willson, Associate Professor
Department of Educational Psychology
Texas A&M University
College Station, TX 77843
409/845-1808

Nancy G. Mangano, Assistant Professor
Kansas State University

William H. Rupley, Associate Professor
Texas A&M University

Sources of Variation that Affect the Reliability of Reading Classroom Observation Measures

Instructional research has a long history of grappling with the development of classroom observation systems that discriminate between effective and ineffective teaching behaviors (Logan, 1981). Reliability of observation systems for use in classroom reading instruction is of particular concern since much of the process/product research utilizes students' reading achievement as a dependent variable. The primary focus of researchers, both historically and with recent reading teacher effectiveness research, in determining reliability of observation systems has been on observer agreement (Scott, 1955; Cohen, 1960; Light, 1971; Frick & Samuel, 1978; Rupley & Mangano, 1982; Rupley, 1982). Minimal attention has been given to either the conceptual analysis of reliability of classroom observations or

specification of sources of variation in classroom observation systems, which affect the reliability of the measure.

Developers and users of reading classroom observation systems that address only inter-rater reliability are failing to account for several sources of variation which can negatively impact the external validity of their findings. First, unreliability contributes to the error term of any statistical analysis. Magnitude of effects are underestimated and the power to detect effects is lowered. Second, unreliability limits statistically the validity of any tests or measures developed from the observations. Third, unreliability has unpredictable effects on the Type I error rates in complex analyses, such as multiple regression and analysis of covariance. Improving the reliability of classroom observation systems is essential to better detect magnitude of effects in reading process/product investigations, especially if one is attempting to specify the process variance associated with students' product behavior. Results of reading teacher effectiveness inquiries conducted by McDonald (1976) attribute approximately 35 percent of students' end of year reading achievement to teacher effects; therefore the magnitude of effect of any single process variable is going to be small.

Sources of Variation

Cronbach (1970) provided a useful framework to consider sources of variation which contribute to the unreliability of any measurement. It is based on a two-by-two table of temporal and generality factors. The generality factor has two levels--specific and general. Specific

effects are those in which the sources of variation are characteristic of the particular measurement instrument being used, while general effects include sources of variation that affect reliability across any instrument. Temporal factors have two levels--temporary and lasting. Temporary effects are those which influence reliability for a short time, during one administration of a test, for example, lasting effects are persistent and affect the reliability of a measure over longer periods of time. The four entries noted in Table 1--lasting-specific, lasting-general, temporary-specific, and temporary-general, are discussed below with examples of procedures to improve reliability through this control.

Insert Table One here

Although some sources of variation are outside the domain of control of the researcher, others can be attended to through careful conceptualization and planning during both the training and implementation phases. The following suggestions have been generated from the literature on observation and the experience of training and working with observers during a two year research project that used the Group Reading Interaction Pattern Observation Instrument (GRIP) (Mangano & Rupley, Note 1). These suggestions include selecting observers, training observers, and attending to specific aspects of actual observation.

Lasting-Specific Sources

Lasting-specific sources of variation in observation systems affect

the reliability of a particular instrument regardless of the conditions of the situation. Examples of such sources of variation include observer's familiarity with the coding system, observer's attitudes toward the observed activities, and differences between training conditions and actual classroom conditions in which observers are asked to function. Such sources can be controlled by researchers employing appropriate criteria for observer selection and developing training procedures that are content specific to the actual observation tasks.

Adequate training of observers is probably one of the most important methods of reducing the variation in observation. It is recommended that the observer be prepared for training through the acquisition of prerequisite knowledge for both the content of observation and the coding of behaviors. This can be accomplished through the use of a variety of training procedures. A major training procedure is a manual with carefully operationally defined concepts that adequately discriminate between categories and presents examples for each of the concepts under observation. Observers can practice discrimination of categories through the use of scripts and audio-tapes prior to the actual training, which facilitates a better understanding of the content (reading, math, classroom management, etc.) and the coding of situational behaviors explicit to that content. This phase can be perceived as a readiness period, where the focus is intended to reduce the variance among individual observers in terms of how they conceptualize the implementation of the observation system in the actual data gathering settings.

Once training begins it is advisable to discuss and illustrate with examples each behavior category and subcategory separately, until

each discrete behavior is learned. In order to master the sequence of behaviors at a faster pace, video-tapes can be utilized. Short segments of simulated actual classroom activities that call for the coding of small portions of the tapes are recommended before the entire instrument is used. Longer scripts taken from actual classroom situations can be video-taped, followed by training in the real classroom. This transition aids the observer in moving from the artificial conditions of training to the on-site condition.

Attitudes of observers toward the observation system should be closely monitored. Observers should be selected who are familiar with both the content of the system, in this instance reading instruction process behaviors, and the conditions in which they system will be utilized, such as self-contained, elementary reading classrooms. Even though such observers may begin their observation enthusiastically, they can eventually lose interest when no reward for their hard work is provided. Paid volunteers are more likely to have a reason for performing proficiently while unpaid volunteers are more apt to become disenchanted or less committed after a period of time. If money is unavailable, course credit, coauthorship, or some other reward to maintain a positive attitude and enhance the value associated with the task will better ensure reliable coding of the behaviors under observation.

Lasting-General Sources

Lasting-general sources include cognitive ability of observers, observation skills, complexity of the observation tasks, and supervision conditions. Selection of observers is often beyond the control of the researcher. For example, in a college environment, observers in

research projects are often gathered from graduate students working under the researchers, students in the classroom who are receiving credit for the observation, and unpaid volunteers. While this is not the most desirable situation, it is realistic. However, when a researcher does have the opportunity to select observers the following suggestions are recommended. First select observers who are capable of mastering the instrument in a relatively short time. Observers who must be trained for longer periods of time and with more individual help throughout the training period will not only affect the time constraints of the training period, but also tend to code less accurately, thus affecting the reliability of the measure. Observers who are too analytical, although helpful in the development stage of an instrument where one wishes to find categories that are ambiguous, will tend to be less reliable in their coding since they can often justify placement of behaviors in more than one category. As noted earlier, observers' lack of a background in the classroom situation that they are observing can affect the reliability of their coding. For example, an observer who has no background in reading instruction may not be aware that certain activities are examples of structural analysis, phonic analysis, or contextual analysis. This may lend to incorrect coding. Equally important are good observation skills and an attitude toward reading that is compatible with the system being used.

Observation system conditions contribute to lasting general variation in the choice of the level of complexity of observation task. Nauroll & Cohen (1973) discuss this in enthnographic field method,

counseling that observers must either concentrate on detail (molecular) or general impressions (molar) but that they should not mix such observations at one time.

Such a recommendation for classroom observation system development and utilization should be carefully addressed. Reliability data gathered during both the training and implementation phases must be appropriate to the data analyses used to explore the significance of results. If reliability coefficients are within an acceptable range for major observation categories, observers' reliability for subcategories for the major categories must also be addressed if these data are to be used in the analyses. Variation within major categories can often be small due to the fact that they are dealing with molar behaviors; however, subcategories often focus on molecular behaviors, which results in increased complexity in accurate specification of behaviors by observers. Although many molecular behaviors are important, for example questioning strategies, attention must be given to defining as accurately as possible during observers' training what constitutes those molecular behaviors. If during observers' training such discrete behaviors can be operationally defined and illustrated; then, the level of inference across and within observers can be minimized.

Another source under control of the experimenter is termed here the supervision condition. Frequent supervision heightens observer awareness and improves consistency of coding. Among the options available to the researcher for observer supervision are: 1) scheduled meetings with individual observers to talk through with them recently completed observation and 2) scheduled classroom observations with

each of the observers. Individual discussion sessions can focus on coding problems specific to a given behavior(s) and/or conditions of observations. Such sessions can often further identify areas of ambiguity across all observers or areas of questionable reliability for individuals. In either case, retraining of all observers for an ambiguous behavior may be warranted. Portions of the system that are unreliable across all observers may be identified for deletion from later data analysis.

Scheduled observations where the researcher codes behavior along with each observer is another means of supervision. Also, data gathered from such paired observations can be used to compute criterion-related reliability coefficients, which address the use of the system by observers in relation to the developer's intended use.

Temporary-Specific Sources

These sources are often not under experimenter control. They are situational and unpredictable. Included are fatigue, attention span, memory of events, practice effect and guessing.

Fatigue involves the time that observers must observe and the periods of times between the periods of time that the observer goes into the classroom and codes behavior. In reading classroom observation one hour is typically sufficient. The closer two observation periods are to each other the more reliable the results. The length of time between periods can affect the observer's memory of categories. If significant time between observations is necessary for the research

project at hand, it is wise to continue practicing through videotape or other means.

Attention span is related to fatigue. The amount of time which the observer can devote to a single event depends on the intensity required. Adult attention spans for nontaxing materials are at least one hour. Intensive activities may have shorter spans, particularly if they are boring or repetitive components. Construction of the coding scheme can help to limit the possibility for loss of attention by keeping individual events short.

Memory problems are minimized by reducing the time between observation and coding. Often this time is instantaneous, but in molar observation field notes must be transcribed while the memory is fresh, usually within 24 hours (Nauroll & Cohen, 1970).

Practice effects involve rehearsal and training on the coding scheme. Certain events may not be recognized when the observer sees them only rarely; therefore, valued but infrequently occurring behaviors about which one wishes to gather data should be reintroduced frequently to maximize observer recognition of them.

Guessing, which occurs when the observer is uncertain as to what occurred, probably cannot be controlled. Behaviors for which observers' use guessing may be identified through the individual supervision sessions that were discussed earlier. It is comforting to assume that the previous training will improve the probability of an accurate coding of the event.

Temporary-General Sources

These sources, general to all observation systems but of limited deviation, are generally amenable to control. Included are physical

location and surroundings, observer drift, and physical condition of the observer. The first consideration is the physical placement of the observer during observation. If the observer cannot hear or see what is going on, he/she might get unreliable results. Yet it is undesirable for the observer to move around during observation because it disrupts the flow of coding and proves to be more obtrusive to the classroom. Careful selection of a place to sit to observe is essential and although seems obvious is often taken for granted until a poor physical placement occurs.

Observer drift refers to the tendency of observers to become more reliable within an observation session from beginning to end, unless fatigue or other factors cause a drop-off in reliability later in the session. This can be handled by a "warm-up" period in which observations are made but not used in later analyses. A comparable situation occurs in door-to-door surveys. The early results from a surveyer are often discarded.

Physical condition of the observer is likely to affect reliability of the observer. Key in importance is awareness of illness during an observation session. If the observer is ill or uncomfortable the session should be cancelled. If the session occurred and the illness reported later the observation are suepect and should not be incorporated in analysis.

Insert Table Two here

Summary

Classroom reading observation systems are complex and maintenance of reliability of measurement is a major concern to the researcher for sorting out the effective process variables and their effects on students' reading achievement. Two dimensions were identified that delineate major sources of variation in observation. The first dimension is generality and the second, temporality. The generality factor has two levels, specific and general, and the temporal factor has two levels, lasting and temporary. The four combinations were discussed with specific examples of each. For each example a suggested method to enhance reliability was provided. The systematic consideration of these sources of variation, that affect reliability is critical to researchers who wish to further our knowledge about classroom interaction and behavior.

Reference Note

1. Mangano, N.G., & Rupley, W.H. Group Reading Interaction Pattern Observation Instrument. Unpublished test, Texas A&M University, 1981.

References

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L.J. Essentials of Psychological Testing. New York: Harper & Row, 1970.

Frick, T. & Semmel, M.I. Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 1978, 48, 157-184.

Light, R.J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365-377.

Logan, J. Classroom observation instruments: An historical perspective. Paper presented at National Reading Conference, Dallas, December, 1981.

McDonald, F.J. Report on phase II of the beginning teacher evaluation study. Journal of Teacher Education, 1976, 27, [illegible]-42.

Nauroll, R. & Cohen, R. A Handbook of Method in Cultural Anthropology. Chicago: Aldine, 1970.

Rupley, W.H. & Mangano, N. Development and measurement issues associated with classroom observation systems. New inquiries in reading research and instruction, (J. Niles and L. Harris, Eds.) Rochester, NY: National Reading Conference, 1982, 200-203.

Rupley, W. Reading teacher effectiveness research: Generalizability of significant findings. Paper presented at National Reading Conference, Dallas, December, 1982.

Scott, W.A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-324.

Table 1: Sources of Variation in Observation Scores in Classroom Observation of Reading

	Lasting	Temporary
Specific	Knowledge of coding scheme	Fatigue due to current task
	Attitude toward activities being observed	Attention span
	Observation constraints	Memory of coding scheme
		Practice effect of coding task
		Luck or guessing
General	Intellectual ability	Observer drift
	Skill in observing	Health, fatigue, emotions
	Observer-wiseness	Motivation
	Attitude toward reading	Physical surroundings
	Molar or molecular level of observation task	Amount of Practice
	Over-analysis	
	Hawthorne effect	

Table 2: Sources of Variation in Observation Systems and Suggested Techniques for Improvement of Reliability

I. Lasting Specific	
A. Knowledge of coding scheme	Training, with testing of knowledge and sequential practice with increasingly complex settings
B. Attitude toward activities being observed	Selection of observers whose philosophies are compatible with the observation system
C. Observational constraints	Train observers to operate in artificial, taped, and unfamiliar conditions.
D. Reading-wiseness	Selection of experienced observers
II. Lasting General	
A. Intellectual ability	Selection of observers with perceptual and cognitive capability to perform the requisite tasks
B. Observational Skill	Thorough training
C. Observation-wiseness	Selection of equally experienced observers or stratification by experience
D. Attitude toward reading	Selection of observers who exhibit objectivity concerning reading theories
E. Molar or molecular level of observation task	Define the level of complexity of conditions to be observed, maintain level during single observational settings; minimize level of inference
F. Hawthorne effect	Maintain supervision, observer awareness of supervision

Table 2 continued

III.	Temporary Specific	
	A. Fatigue due to current task	Limit time of observation for the task
	B. Attention span	Limit time for observation of any category, event or time
	C. Memory of event	Minimize time between observation and coding
	D. Memory of the coding scheme	Limit complexity of the scheme, maintain practice or rehearsal schedule
	E. Practice effect of current task	Maintain practice or rehearsal schedule close in time to actual observation
	F. Luck or guessing	——
IV.	Temporary General	
	A. Observer drift	Segment observations by time within observation sessions and across observation sessions; analyze separately and combine only if no evidence of drift
	B. Health, fatigue, emotions	Disallow observations made under extremes of any of these conditions in observers
	C. Motivation	Use internal or external motivators to promote good observation
	D. Physical surroundings	Arrange for good viewing and hearing
	E. Practice	Provide for continuous practice in actual or artificial settings, use live or taped media