DOCUMENT RESUME

ED 093 944 TM 003 762

AUTHOR	Stallings, Jane; Giesen, Phillip A.
TITLE	A Study of Confusability of Codes in Observational Measurement.
INSTITUTION	Stanford Research Inst., Menlo Park, Calif.
PUB DATE	[Apr 74]
NOTE	50p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th Chicago, Illinois, April 1974)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Bias; *Classroom Observation Techniques; Codification; Measurement Techniques; *Reliability; Simulation; *Video Tape Recordings

ABSTRACT

Classroom observation has the potential of obtaining valuable information regarding teacher and child behavior. This study examines the accuracy of observers coding a standard stimuli. In this procedure the observer's bias is examined, as well as the confidence that can be placed in the observation code itself. Through these procedures the exact nature of the confusion of codes can be identified. In order to avoid the problems encountered with the paired observer method, an attempt was made to assess the accuracy of observers through the use of controlled videotape examples which allow each interaction (a frame) and sequences of frames to be analyzed for accuracy. Ten videotaped skits were produced to present concise, clear examples of each code used in recording classroom interactions on an observation instrument. Confusability (low observer agreement) matrices were constructed by tallying the observer code sequences. Results of the confusability study identify the specific codes that appear to be reliable as well as those that are confused and need to be redefined. Inter-rater accuracy and videotape simulation accuracy are compared. While the two systems of examining observer accuracy do yield some different information, it is not contradictory, and the videotape system is easier to interpret. (Author/RC)

STANFORD RESEARCH INSTITUTE
Menlo Park, California 94025 · U.S.A.

A STUDY OF CONFUSABILITY OF CODES IN OBSERVATIONAL MEASUREMENT

by

Jane Stallings, Ph.D.
Phillip A. Giesen

Presented to the
American Educational Research Association
1974 Annual Meeting
Chicago, Illinois

April 15-19, 1974

Table of Contents

Introduction 1

A. Procedure to Assess the Confusability of Observation Codes . . . 2

1. A Description of Procedures 2

2. Procedural Problems 3

B. A Description of Confusability Matrices 4

C. Confusion of Observation Codes (A Study Combining the Results of 63 Observers) 11

1. Findings of "What" Code Confusions 15

2. Findings for "How" Code Confusions 18

3. Summary 20

D. Accuracy of Individual Observers 20

1. Findings from "What" Codes Occurring Six or More Times . . 23

2. Findings from "How" Codes Occurring Six or More Times . . . 24

3. Summary 24

E. A Study Comparing Inter-Rater Accuracy and Videotape Simulation Accuracy 25

1. Paired Observers 25

2. Videotaped Skits (Simulation) 30

3. Conclusions 36

A STUDY OF CONFUSABILITY OF CODES IN OBSERVATIONAL MEASUREMENT

As a measurement technique, classroom observation has the potential of obtaining valuable information regarding teacher and child behavior. This potential is realized to the extent that the observation data can be shown to be reliable. Reliability, as Cronbach et al (1972) have pointed out, is related to the notion of generalizability from a sample to some universe of interest.

Several sources of unreliability have been identified in past research. Medley & Mitzel (1963) say that:

> Most commonly, it [unreliability] occurs when two measures of the same class tend to differ too much; this may happen because the behaviors are unstable, because the observers are unable to agree on what occurs, because the different items which enter into the measurement lack consistency, or for some other reason.

Neither Cronbach nor Medley address the question of the confusability of codes used in observation instruments. The present study examines the accuracy of observers coding a standard stimuli. In this procedure the observer's bias is examined, as well as the confidence that can be placed in the observation code itself. Through these procedures the exact nature of the confusion of codes can be identified.

In previous SRI reliability studies, the technique of pairing the observers with an SRI trainer has been used. However, there are some problems in assessing inter-rater reliability. First, there is some variability in the coding skills of SRI trainers. Second, there is most certainly a variability in the incidents which occur in the classrooms, in what is selected for observation, and in which codes are used in the observations. The optimum arrangement might be to have all observers and SRI trainers observe the same phenomena in the same classroom at the same time. But, as Soar (1973) says:

> The critical problem (of paired observers) is the effect on the classroom of increasing the number of observers. One observer represents a threat to many

teachers and a distraction to the children, at least initially, and as the number of observers increases, these difficulties increase, probably more like a geometric function than an arithmetic one.

In an effort to avoid the problems encountered with the paired observer method, SRI staff has attempted to assess the accuracy of observers through the use of controlled videotape examples. This procedure allows each interaction (or frame) and sequences of frames to be analyzed for accuracy, whereas previously only simple marginal frequency counts of single codes could be computed.

Other investigators in observational research also use videotapes to assess observer accuracy. Soar (1973) used tapes of actual classroom events, and Simmel (1973) cleverly used the last ten minutes of the Johnny Carson Show to check observer accuracy on a weekly basis. Although they are useful, the limitations of videotapes also should be recognized:

- Because of the difficulty in seeing and hearing, videotapes are more difficult to code than live conversations;
- It is more difficult to understand the gestalt of the situation from a tape than it is from a live situation in the classroom;
- Simulated skits are likely to be more clear-cut examples than those which actually occur in classrooms.

A. PROCEDURE TO ASSESS THE CONFUSABILITY OF OBSERVATION CODES

1. A Description of Procedures*

Differing from both Soar and Simmel, SRI staff produced ten videotaped skits. Each simulation is approximately 20 interaction frames long. These skits attempt to present concise, clear examples of each code used in recording classroom interactions on the SRI observation instrument. Each skit begins with a still picture and the voice of a narrator who explains the situation and identifies the focus person. The

*These procedures were developed at SRI by J. Philip Baker, Phillip Giesen, and Charles Norwood.

skit is then shown at regular speed. After the skit is shown once, the still picture and narrator again identify the focus person. Each skit is then shown again, this time with a 2- to 3-second pause between each interaction. The observers are instructed to code this stop-action portion of the skit and to code one frame during each stop or pause.

2. Procedural Problems

The reliability coding booklets were returned to SRI and compared with the criterion sequences. This revealed that some observers were coding more than one frame during a pause. Conversely, some observers, possibly while turning pages, omitted frames. The trainers reviewed the coding sequences and deleted extraneous frames or inserted spaces so as to align the observers' sequences with the criterion sequences. Three trainers performed this operation. Since judgment is involved, a check was made on the code sequences of 10 observers to see whether the trainers arranged the sequences in the same manner. The average agreement between trainers in arranging these sequences was 96.4 percent.

Other procedural problems were also encountered due to the experimental nature of the techniques used. Comments received from the observers indicated that not all of the equipment utilized to administer the tapes was in good condition, and, as a result, the sound or pictures were of poor quality. Also, some examples on the criterion tape were technically less than well executed. The most serious problem, however, was that there were too few examples of several of the codes on the criterion tape. Five or fewer examples of a code limited the assurance that representative examples of the code were shown. Further, if an observer missed two out of four possible instances of a code, he only had a score of 50 percent of the criterion correct; however, if he missed two out of 30 possibilities, he had a score of 93 percent of the criterion correct. For this reason, the codes which have fewer than six examples will not be interpreted in this analysis. The number of examples of each code on the complete set of tapes ranges from zero to 40. (This problem is being remedied by the development of more skits.)

B. A DESCRIPTION OF CONFUSABILITY MATRICES

Confusability of codes refers to codes which were confused with the correct codes by an observer.* Confusability matrices were constructed by tallying the observer code sequences. For each frame, a tally mark was entered in the box or cell created by the juncture of the criterion code and the code marked by tbe observer. Figure 1 shows an example of a confusability matrix for the "What" codes. The principal diagonal contains the cells indicating correct coding; other cells contain incorrect coding. The column totals are the total number of criterion examples shown on the videotape for each code; the row totals are the total number of times an observer recorded each code, whether correctly or not. An examination of a particular cell reveals whether the code was recorded correctly or incorrectly and, if recorded incorrectly, shows exactly which codes were confused.

The total number of tallies in each cell can be used to calculate the rates of accuracy in two related but distinct ways. The first procedure described above allows an examination of observer bias. If the number in a given cell is compared to the total number of recordings (row total) of the code that pertains to that cell (see the row indicator), a ratio of correct or incorrect responses can be derived. For the cells that fall on the main diagonal, the numbers indicate the proportion of times the code recorded by an observer was correct. For the cells that do not fall on the diagonal, the number indicates the proportion of error. This accuracy rate is called the Accuracy Rate of each observer on each code; i.e., the ratio of correct or incorrect codes of the total number of coded observations.

The point of the second procedure is to assess the confidence that can be placed in each code. The second accuracy figure can be arrived at by comparing the number of tallies located in the same cell to the total number of examples on the criterion tape presented of that specific code (see column total). Again, the number arrived at shows the proportion of correctness to incorrectness, as based on whether the cell falls on the diagonal. This proportion of times the criterion instances were recorded correctly or incorrectly is called the Criterion Accuracy Rate.

*See Table 1 for a brief explanation of the SRI "What" and "How" codes.

Table 1

SRI "WHAT" AND "HOW" CODES

"What" Codes	"How" Codes
1 - Command or Request	H - Happy
1Q - Direct Question	U - Unhappy
2 - Open-ended Question	N - Negative
3 - Response	T - Touch
4 - Instruction, Explanation	Q - Question
5 - Comments, Greetings; General Action	G - Guide/Reason
6 - Task-related Statement	P - Punish
7 - Acknowledge	O - Object
8 - Praise	W - Worth
9 - Corrective Feedback	DP - Dramatic Play/ Pretending
10 - No Response	A - Academic
11 - Waiting	B - Behavior
12 - Observing, Listening	
NV - Nonverbal	
X - Movement	

Figure 1

TOTAL NUMBER OF "WHAT" CODES RECORDED IN EACH CELL BY ONE OBSERVER

Code as Recorded by Observer (by Row) "What" Codes	Criterion Examples (by Column) 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	Total No. Recorded by Observer
1	3																3
1Q		20														1	21
2		1	5														6
3		1		26							1						28
3NV					2												2
4						4											4
4NV							8										8
5								0	1	1							2
5NV									6								6
6			1							16							17
7											8						8
8											1	2					3
9													8				10
10														0			0
11															3		3
12																22	26
Total Number of Examples on Criterion Tape	3	22	6	27	2	7	8	1	7	17	11	2	8	0	3	23	147

Figure 2 presents the proportion of cell tallies to the row totals for each cell. This provides a matrix that presents the Accuracy Rates in place of the raw scores shown on Figure 1. For example, the observer recorded a "1" code correctly three times, so the Accuracy Rate is 100, or 1.00. In the second row, "1Q" was recorded correctly 20 out of 21 times, or 95 percent of the total number of times.

Looking across the "1Q" row, we see that the observer called an example of a "12" a "1Q" five percent of the time. For another example, we look across the row of code "5" and see that our observer did not code any "5's" correctly. Instead, she mistakenly coded two examples of code "5" as a "5NV" and a "6."

In Figure 3, the proportion of criterion examples for each code correctly recorded by the observer are found in the diagonal cells. Entries in cells down the column marked by the correct code other than in the diagonal cells are instances of confusion. As can be seen in the "1Q" column of Figure 3, code "2" and code "[illegible]" were sometimes confused with the "1Q" code. The bottom row of the figure presents the correct number of criterion examples of each code which appeared on the videotape. The last column on the table lists the number of times the observer recorded each code. In the code "12" column on Figure 3, the observer recorded three more "12's" than appeared on the videotapes. Apparently, these three were confused with some other code.

Figures 2 and 3 can be overlaid so that the top entry in a cell refers to the accuracy of what the observer recorded, and the lower entry refers to the percent of criterion examples which were correctly coded. Figure 4 illustrates such an overlay. The combined figure tells us that when this observer recorded a "7," it was indeed a "7" (there are no other entries in the code "7" row). However, she only recorded eight "7's" and there were actually 11 examples on the videotapes. Looking down the column for code "7" and at the lower entry in the cell, we see that examples of "7" were recorded as "3" nine percent of the time, as "8" nine percent of the time, and as "12" nine percent of the time. We can conclude from this example that when the observer recorded a "7" it was truly a "7," but she underestimated the number of times they occurred. She recorded some of the "7's" as "3," "8," or "12."

Figure 2

PROPORTIONAL DISTRIBUTION OF CODES AS RECORDED BY ONE OBSERVER (The Rows)

"What" Codes	Criterion Examples: 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	Total No. Recorded by Observer
1	1.00																3
1Q		.95														.05	21
2		.17	.83														6
3		.04		.93							.04						28
3NV					1.00												2
4						1.00											4
4NV							1.00										8
5								0	.50	.50							2
5NV									1.00								6
6			.06							.94							17
7											1.00						8
8											.33	.67					3
9				.10				.10					.80				10
10														0			0
11															1.00		3
12						.12					.04					.85	26
Total Number of Examples on Criterion Tape	3	22	6	27	2	7	8	1	7	17	11	2	8	0	3	23	147

Code as Recorded by Observer (by Row)

Figure 3

PROPORTIONAL DISTRIBUTION SHOWING HOW CRITERION EXAMPLES WERE RECORDED BY ONE OBSERVER (The Columns)

"What" Codes (Criterion Examples by Column)	1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	Total No. Recorded by Observer
1	1.00																3
1Q		.91															21
2		.05	.83														6
3		.05		.96							.09						28
3NV					1.00												2
4						.57											4
4NV							1.00										8
5								0	.14	.06							2
5NV									.86								6
6										.94							17
7											.73						8
8											.09	1.00					3
9													1.00				10
10														0			0
11															1.00		3
12																.95	26
Total Number of Examples on Criterion Tape	3	22	6	27	2	7	8	1	7	17	11	2	8	0	3	23	147

Codes

Figure 4

PROPORTIONAL DISTRIBUTION OF "WHAT" CODES FOR ONE OBSERVER FOR BOTH RECORDED AND CRITERION ACCURACY

Criterion Examples (Lower Numbers in Cells)

"What" Codes	1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	Total No. Recorded by Observer
1	1.00 1.00																3
1Q		.95 .91														.05 .04	21
2		.17 .05	.83 .83														6
3		.04 .05		.93 .96							.04 .09						28
3NV					1.00 1.00												2
4						1.00 .57											4
4NV							1.00 1.00										8
5								0 0	.50 .14	.50 .06							2
5NV									1.00 .86								6
6			.06 .17							.94 .94							17
7											1.00 .73						8
8											.33 .09	.67 1.00					3
9				.10 .04				.10 1.00					.80 1.00				10
10														0 0			0
11															1.00 1.00		3
12						.12 .43					.04 .09					.85 .96	26
Total Number of Examples on Criterion Tape	3	22	6	27	2	7	8	1	7	17	11	2	8	0	3	23	147

Codes as Recorded by Observer (Upper Numbers in Cells by Row)

C. CONFUSION OF OBSERVATION CODES (A STUDY COMBINING THE RESULTS OF 63 OBSERVERS)

Analyses of these matrices were used in two ways. First, by combining the results for all observers, the extent of general confusability of codes could be examined. Codes that reveal a high rate of confusion by several observers suggest these possible causes: an overlapping of code definitions, poor videotape examples, or less-than-adequate training procedures. Second, the accuracy of individual observers could be examined with these matrices. For example, if an observer were not very accurate on code "8" (praise), then codes using "praise" could be examined for anomalies.* The findings reported here represent 63 observers spread among 30 geographical locations.

How the observers coded the videotaped examples is shown in Figure 5. The diagonal shows the number of correct codings. The row at the bottom of the table is the number of videotape criterion examples coded by all of the observers. Each figure in the bottom row can be compared with the corresponding cell in the diagonal. For example, code "1" was recorded correctly 160 times out of a possible 245 times. The other entries in the "1" column are sources of confusion.

The proportion of times that the observers were correct in their recordings is presented in Figure 6. For example, the observers recorded "1Q" correctly 80 percent of the time. Looking across the "1Q" row, we see that two percent of the time when a "1Q" was recorded it was truly a "1," and 11 percent of the time it was truly a "2."**

The proportion of videotaped examples recorded by the 63 observers is shown in Figure 7. The number in the diagonal reports the percent recorded correctly. The numbers in the columns outside of the diagonal indicate the source of error. If all of the numbers in the columns were in the diagonal cell, the result would be 100 percent correct. The total number of possible examples on the videotape are listed on the bottom row. For example, "1Q" was recorded correctly 77 percent of the time, whereas

---

*Caution: Even if the observer were 100 percent in agreement with the criterion examples, in a study of this type generalization would still be limited by the day-to-day variability of classroom events.

**Sources of error less than three percent are not included on the table.

Figure 5

TOTAL OF "WHAT" CODES RECORDED IN EACH CELL BY 63 OBSERVERS

"What" Codes	Criterion Examples: 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12	Total No. Recorded by Observer
Codes as Recorded by Observers: 1	160	13	2	5	1	3	0	3	0	9	1	0	13	0	0	12	222
1Q	27	966	135	24	1	10	0	0	0	4	4	0	29	0	0	12	1212
2	0	153	156	3	1	0	0	0	0	0	1	0	3	0	0	0	317
3	1	35	4	1310	17	2	0	4	2	29	14	0	10	0	0	19	1447
3NV	2	0	0	14	142	2	0	0	9	4	4	0	0	0	0	3	180
4	2	9	0	34	1	344	11	0	1	7	8	0	8	0	0	55	480
4NV	2	1	0	5	2	3	480	0	31	2	13	2	0	0	2	5	548
5	14	7	0	18	1	2	19	37	39	200	3	3	16	0	0	6	365
5NV	1	1	0	0	3	2	96	2	343	6	5	2	0	0	17	1	479
6	27	31	22	75	1	21	11	5	3	581	36	0	29	0	0	11	853
7	0	3	1	38	1	7	0	0	0	1	455	5	10	0	0	6	527
8	0	0	0	7	0	1	0	0	0	4	22	106	2	0	0	3	145
9	3	3	0	9	0	13	2	0	23	2	1	378	0	0	4	439
10	1	3	0	1	11	1	0	0	6	0	0	0	0	0	3	5	31
11	0	0	0	1	0	0	0	0	5	0	0	0	0	0	125	1	132
12	5	27	0	28	6	62	4	3	19	11	37	2	4	0	29	1190	1427
Number of Examples over All Observers	245*	1252	320	1572	188	473	622	56	458	881	605	121	502	0	176	1333	8804

Figure 6

PROPORTIONAL DISTRIBUTION OF CODES RECORDED BY 63 OBSERVERS

Codes as Recorded by all Observers	Criterion Examples: "What" Codes 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12
1	.72*	.06											.06			
1Q	.02	.80	.11													
2		.48	.48													
3		.02		.91						.02						
3NV				.08	.79											
4				.07		.72										.11
4NV							.88		.06							
5				.05			.05	.10	.11	.55			.04			
5NV							.20		.72							
6	.03		.03	.09						.68	.04		.03			
7				.07							.86					
8											.15	.73				
9										.05			.86			
10														0		
11															.95	
12																.83
Total Number of Examples on Criterion Tape	4	20	5	26	3	8	11	1	7	16	11	2	9	1	3	23

*Only numerals in the diagonal report proportion correctly coded.

Figure 7

PROPORTIONAL DISTRIBUTION SHOWING CRITERION EXAMPLES RECORDED BY 63 OBSERVERS**

"What" Codes	Criterion Examples: 1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12
1	.65*	.01											.03			
1Q	.11	.77	.42										.06			
2		.12	.49													
3		.03		.83												
3NV					.76											
4						.73										.04
4NV							.77									
5								.66								
5NV							.15		.75						.10	
6	.11		.07	.05	.09	.04		.09		.66	.06		.06			
7											.75	.05				
8												.88				
9													.75			
10														0		
11															.71	
12						.13					.06				.16	.89
Total Number of Examples on Criterion Tape	4	20	5	26	3	8	11	1	7	16	11	2	9	1	3	23

"Codes"

16 percent of the time it was confused with the codes "1," "2," and "3."

Figure 8 is an overlay of Figures 6 and 7. This provides the data necessary to quickly assess the observers' accuracy (by looking at the topmost entry in a cell) and the percent of criterion codes which have been recorded (by looking at the lower entry in a cell).

1. Findings of "What" Code Confusions

Since a "What" code is required for each interaction, each recorded frame must include a recording of a "What" code. The observer only has the option of recording the correct (or criterion) code or recording the wrong code. The entire frame is considered void if no "What" code is recorded.

In Figure 8, four of the "What" codes have been separated into two categories: the "What" code alone and the "What" code with its "How" modifier. This was done because the meaning or definition of the "What" code is modified or sometimes changed by the addition of these specific "How" codes. An example of this is the "5" code. The definition of the "5" is "general comment," but the definition of the "5NV" is "general action."

As mentioned earlier, the number of criterion examples for some codes is small which limits the conclusions that can be drawn regarding these low frequency codes. For this reason, codes with fewer than six examples on the videotapes will not be discussed.

Nine of the 16 "What" codes have six or more criterion examples of each code. These are the shaded diagonal cells in Figure 8. Of these nine ("1Q," direct question; "3," response; "4," instruction; "4NV," self learning, "5NV," general action; "6," task-related comment; "7," acknowledge; "9," corrective feedback' and "12," attending), only "6" has an observer accuracy rate that is lower than .70.

Code "6," task-related comment, was confused most often with code "3," response. It was also sometimes confused with examples which were actually "1," direct request, "2," open-ended questions, and "9," corrective feedback (see row "6"). This suggests that the definitions and training procedures need to be more exact regarding when to code a task-related comment "6." The numbers in the lower section of the cell

Figure 8

PROPORTION OF "WHAT" CODES RECORDED IN EACH CELL, CALCULATED OVER 63 OBSERVERS*

"What" Codes as Recorded by Observers (Upper Numbers in Cells by Row)	Criterion Examples (Lower Numbers in Cells)															
	1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	10	11	12
1	.72 .65	.06 .01											.06 .03			
1Q	.02 .11	[illegible]	.11 .42										.06			
2		.48 .12	.49 .49													
3		.02 .03		[illegible]						.02						
3NV				.08	.79 .76											
4				.07		[illegible]										.11 .04
4NV							[illegible]		.06 .07							
5				.05			.05	.10 .66	.11 .09	.55 .23			.04			
5NV							.20 .15		[illegible]						.10	
6	.03 .11		.03 .07	.09 .05	.09	.04		.09		[illegible]	.04 .06		.03 .06			
7				.07							[illegible]	.05				
8											.15	.73 .88				
9										.05			[illegible]			
10														0		
11															.95 .71	
12						.04 .13					.06				.16	[illegible]	
Total Number of Examples on Criterion Tape	4	20	5	26	3	8	11	1	7	16	11	2	9	1	3	23

*Sources of error less than three percent are not included on this table.

(looking down the "6" column) indicate that 23 percent of the time the criterion examples of "6" were recorded as "5."

The next lowest in reliability was the "4" code, instructing. Eleven percent of the time the observers recorded what was actually "12," observing, as a "4." Since "4" is verbal and "12" is nonverbal, the problem would not appear to be one of confusion in the normal sense but, rather, confusion of which person to focus upon. This conclusion is based on the fact that both of these codes generally occur simultaneously (that is, when a teacher is instructing, "4," the children are usually attending, "12"). Apparently the observers confused which person to record. As can be seen in Figure 8 in the code "12" row, a true example of "4" was sometimes confused and recorded as a "12," which is a further indication that the instructions regarding the focus of observation were not clearly understood by observers.

Code "4NV" describes a child working alone on instructing himself. Observers recorded this reliably 88 percent of the time. They sometimes confused "4NV" with what was truly a "5NV," a code that describes "play" rather than "self instruction in a task." Looking down the "4NV" column, it can be seen that 15 percent of the videotaped examples were recorded as "5NV's." This confusion of "4NV" and "5NV" indicates an overlap of definitions (or a conceptual difficulty in distinguishing "work" from "play").

Criterion examples of code "7," acknowledgement, were sometimes confused with "6" and "12." Code "7" is sometimes confused with code "3," responding (see row 7 in Figure 8). It is easy to see how acknowledging a child can be confused with responding to a child. On the other hand, code "3," responding, was one of the more reliable codes. It was not confused with "7" (see Figure 8). In fact, the observers recorded it correctly 91 percent of the time, and column 3 indicates that five percent of the criterion examples were confused with "6."

Eleven percent of the recorded code "1Q's," asking direct questions, were actually code "2's," asking open-ended questions. The confusion between "1Q" and "2" has long been recognized by the SRI researchers. Each year the variables have been defined more carefully; however, there still seems to be a gray area of unclarity between the two

codes. Code "2," which has too few examples to analyze with confidence, was also confused with "1Q." The results of individual observers were examined, and apparently those observers who observed models which do not often require the "2" code had a higher rate of error.

Eighty-six percent of the time the observers recorded "9," corrective feedback, correctly; five percent of the time code "6" was recorded as "9" (see row 9, the upper value). The criterion examples as illustrated in column 9 (the lower value) were sometimes recorded as "1," "1Q," and "6."

2. Findings for "How" Code Confusions

A "How" code is not always required. This rule leads to four distinct possibilities:

- A required "How" was left out of the frame (omission). These are listed at the bottom of Figure 9;
- A "How" code was recorded when not called for (intrusion). These are listed in the last column of Figure 9;
- The criterion "How" code was confused with another code. These are entered in other than the diagonal cells;
- The criterion "How" code was recorded accurately. These are entered in the diagonal cells.

Only six of the 14 "How" codes were represented by six or more examples on the videotapes. These are "NV," "X," "A," "B," "DP," and "O" (see Figure 9). As described on page 3, codes with fewer than six examples will not be discussed. Also as described earlier, the upper value in a cell reports the percent of observer accuracy. The lower value in the cell reports the percent of the videotaped examples which were correctly recorded.

The nonverbal code "NV" was recorded correctly 93 percent of the time by observers; and, overall, the observers omitted only 13 percent of the criterion examples. Code "X," movement, was also found to be reasonably reliable. Eighty-nine percent of the time the observer recorded it correctly, but 20 percent of the examples were om ed by observers.

Figure 9

PROPORTION OF "HOW" CODES RECORDED IN EACH CELL, CALCULATED OVER 63 OBSERVERS

Criterion Examples (Lower Numbers in Cells)

"How" Codes	NV	X	A	B	DP	W	O	P	G	Q	T	N	U	H	Proportion of Intrusions
NV	.93 .86														.07
X		.89 .81													.10
A			.81 .76	.04 .13											.15
B				.95 .43											.05
DP					.84 .55										.15
W						.79 .61									.16
O							.81 .66								.18
P								0							0
G									.13 .33	.13 .66		.20 .19			.51
Q										.87 .75					.10
T											.68 .68				.31
N												.40 .43	.03		.53
U												.03	.70 .40	.03	.23
H														.63 .45	.35
Proportion of Omissions	.13	.20	.21	.43	.43	.39	.33	0	.66	.10	.32	.33	.56	.52	
Total Number of Examples on Criterion Tapes	26	6	40	13	9	2	8	0	1	2	3	3	4	2	

Codes as Recorded by Observers (Upper Numbers in Cells by Row)

Observers recorded the "A" code, academic, correctly 81 percent of the time. Four percent of the "A's" recorded were truly code "B," and 15 percent of the "A's" were actually intrusions. Seventy-six percent of the videotaped examples were recorded correctly, and 21 percent were omitted.

While 95 percent of the examples recorded as code "B" by the observer were correct (see row B), 43 percent of the "B's" were omitted and 13 percent of the examples of "B's" were incorrectly recorded as "A's." This leads to the conclusion that if a "B" is recorded, it is likely to be correct, but the total number of "B" codes may be underestimated by over 50 percent. An examination of each observer's work is important in order to discover the source of the underestimation. It is possible that only a few observers are grossly underestimating "B's," or it could be that many of the 63 observers are underestimating "B's" to only a small degree.

The two remaining codes with six or more examples ("DP," dramatic play, and "O," use of objects) were recorded accurately over 30 percent of the time, but both codes were underestimated (43 percent and 33 percent of the time).

3. Summary

The results of the confusability study identify the specific codes that appear to be reliable as well as those that are confused and need to be redefined. The findings suggest that some codes, such as "6," "4NV," and "5NV," should be more carefully defined because of overlapping definitions. There is some indication that there should be more careful training of observers on the focus of observation so that "4" and "12" will not be confused. The overall reliability for all observers on the "What" codes was 78 percent and 81 percent for the "How" codes.

D. ACCURACY OF INDIVIDUAL OBSERVERS

The value of this new method for measuring accuracy is that it contributes directly toward interpreting the data. Observer bias can be assessed by examining the overuse, underuse, or confusion of codes. In this study, each observer was responsible for observing one grade level

at a single site. Therefore, the data collected by each observer is identifiable in the analysis.

In order to determine the accuracy rates for each observer separately, tables were constructed that graphically present, by sponsor, each observer's results (see Table 2). Thus, for example, if an observer in Grade 1 at Site X had difficulty with the code "7," acknowledgment, it is possible to compute the site mean of code "7" and compare it with the first grade means of code "7" at the four other sites of the sponsor. If the means of the four sites (not in question) are similar and the mean of the site in question differs from the other four, there are two possible explanations: (1) Site X may be truly different from the other four sites, or (2) the observer at Site X may not be recording accurately. In any case, the data resulting from code "7" at Site X would be interpreted with caution. This procedure allows for each observer's data to be reviewed in order to estimate the accuracy of the individual on each code and to allow for the data to be interpreted accordingly.

As an example, Table 2 shows the observer accuracy rate (the top number) and the criterion accuracy rate (the bottom number) for each of the Far West observers for each code. In addition, an overall accuracy rate for each observer on all "What" and "How" codes has been computed and displayed on this table to provide a general idea of the observer's skill.* The results are grouped by grade level and site. Similar tables for the other six sponsors in the evaluation were prepared. The complete confusability matrix of all observers is not included in this report but is available at SRI.

As previously discussed on page 3, five or fewer criterion examples of a code minimize the confidence with which the actual results can be

*The overall accuracy rate is arrived at by computing the ratio of correct recordings (those that fall in the diagonal cells) of all codes to the total number of recorded codes and to the total number of criterion instances of the codes. For the "What" codes, the two ratios are the same since the total number of recorded codes is equal to the total number of criterion instances. Two ratios are required for the "How" codes since observers are not required to record a "How" code in each frame which leads to differences between the total numbers of criterion examples and total numbers of recorded codes.

Table 2

ACCURACY RATES* FOR THE "WHAT" AND "HOW" CODES BY FAR WEST OBSERVERS

FW SPONSOR	"WHAT" CODES — CRITERION INSTANCES — SIX OR MORE: 10	3	4	4NV	5NV	6	7	9	12	"WHAT" CODES — CRITERION INSTANCES — FIVE OR LESS: 1	2	3NV	5	8	11	"HOW" CODES — CRITERION INSTANCES — SIX OR MORE: NV	X	A	B	DP	O	"HOW" CODES — CRITERION INSTANCES — FIVE OR LESS: W	G	Q	T	N	U	H	OVERALL ACCURACY RATE: "WHAT"	"HOW"
0201 Berkeley																															
Grade 1	.84/ .80	.96/ .89	.62/ .89	.82/ .90	1.0/ .50	.93/ 1.0	.90/ .82	.86/ .75	.94/ .81	1.0/ 1.0	.57/ .80	.50/ 1.0	0.0/ 0.0	.67/ 1.0	1.0/ 1.0	1.0/ .85	1.0/ .80	.67/ .90	1.0/ .70	1.0/ .67	1.0/ .67	.67/ 1.0	0.0/ 0.0	1.0/ .50	1.0/ .50	.14/ .50	1.0/ .50	.50/ 1.0	.84	.74/ .78
Grade 3	.81/ .81	1.0/ .96	1.0/ .38	1.0/ .46	.50/ .88	1.0/ .81	.90/ .82	.73/ 1.0	.82/ .96	.05/ .25	.60/ .60	.75/ 1.0	.50/ 1.0	.67/ 1.0	1.0/ 1.0	.89/ .96	.83/ .83	.82/ .78	.88/ .54	1.0/ .78	1.0/ .25	.67/ 1.0	1.0/ 1.0	.50/ .50	1.0/ .67	.50/ .50	.50/ .25	1.0/ .67	.83	.83/ .73
0204 Duluth																															
Grade 1	.86/ .70	.96/ .84	.86/ .75	.83/ .91	.71/ .71	.62/ .47	.86/ .75	.71/ .63	.92/ .96	.75/ .75	.60/ .60	1.0/ 1.0	.09/ 1.0	.50/ 1.0	1.0/ 1.0	.93/ .96	.67/ 1.0	.66/ 1.0	1.0/ .50	.88/ 1.0	.88/ 1.0	.50/ .50	0.0/ 0.0	1.0/ .50	.75/ 1.0	0.0/ 0.0	1.0/ .25	.50/ .50	.78	.74/ .85
Grade 3	.67/ .80	.79/ .76	.67/ .67	.67/ .67	.80/ .67	.33/ .29	.60/ .67	.71/ .63	.94/ .80	.50/ .50	.43/ .60	.50/ .50	.20/ 1.0	.50/ .50	1.0/ .67	.83/ .64	.83/ .63	.66/ .84	1.0/ .50	0.0/ 0.0	0.0/ 0.0	.50/ 1.0	0.0/ 0.0	0.0/ 0.0	1.0/ .67	.25/ .33	0.0/ 0.0	.20/ .50	.64	.67/ .56
0207 Lebanon																															
Grade 1	.92/ .57	.81/ 1.0	.71/ .83	1.0/ .82	.75/ .86	.77/ .77	.88/ .70	.83/ .83	.91/ .96	1.0/ .25	.36/ .00	.80/ 1.0	0.0/ 0.0	1.0/ 1.0	1.0/ 1.0	1.0/ .96	1.0/ .67	.89/ .69	0.0/ 0.0	0.0/ 0.0	.88/ .88	1.0/ .50	0.0/ 0.0	1.0/ 1.0	1.0/ .33	1.0/ 1.0	1.0/ 1.0	1.0/ .50	.81	.92/ .64
Grade 3	.94/ .81	.96/ .81	.78/ .88	.79/ 1.0	.86/ .75	.75/ .69	1.0/ .70	.88/ .88	.87/ .91	1.0/ .67	.50/ .60	1.0/ 1.0	.14/ 1.0	.67/ 1.0	1.0/ 1.0	.96/ .93	1.0/ .83	.89/ .67	.92/ .69	1.0/ .71	.88/ .88	1.0/ 1.0	0.0/ 0.0	1.0/ .50	.50/ .67	.33/ .75	1.0/ .50	1.0/ .50	.83	.88/ .76
0209 Salt Lake City																															
Grade 1	.80/ .76	.92/ .92	.83/ .63	1.0/ .70	.44/ .57	1.0/ .29	.90/ .82	.67/ 1.0	.75/ .91	.75/ .75	.50/ .40	.75/ .75	.10/ 1.0	.50/ 1.0	1.0/ .33	.84/ .91	.80/ .80	.82/ .97	.86/ .46	.75/ 1.0	.75/ .43	.17/ 1.0	.25/ 1.0	1.0/ 1.0	.67/ .67	.38/ 1.0	.33/ .33	0.0/ 0.0	.74	.71/ .81
Grade 3	.78/ .70	1.0/ .89	1.0/ .71	.92/ 1.0	1.0/ .83	.75/ .56	.90/ .90	1.0/ .63	.79/ 1.0	1.0/ 1.0	.38/ .60	1.0/ 1.0	.13/ .33	.67/ 1.0	1.0/ .50	1.0/ .96	1.0/ 1.0	.70/ .97	.33/ .08	.78/ .88	1.0/ 1.0	.50/ 1.0	.50/ 1.0	1.0/ 1.0	.67/ .67	.33/ .33	.25/ .33	.35/ .50	.81	.75/ .81
0213 Tacoma																															
Grade 1	.75/ .79	.95/ .80	.54/ .78	1.0/ .73	.89/ 1.0	.79/ .94	.86/ .60	1.0/ .89	.76/ .96	.60/ .75	.43/ .60	0.0/ 0.0	0.0/ 0.0	1.0/ 1.0	0.0/ 0.0	1.0/ .85	1.0/ 1.0	.90/ .77	1.0/ .31	.67/ .67	.22/ .25	1.0/ 1.0	0.0/ 0.0	1.0/ 1.0	1.0/ .67	.50/ .33	1.0/ .50	1.0/ .50	.79	.82/ .67
Grade 3	.95/ .73	.92/ .73	.75/ 1.0	.88/ .70	1.0/ .50	.58/ .82	.73/ .80	1.0/ 1.0	.93/ .93	1.0/ .50	.36/ .80	0.0/ 0.0	.25/ 1.0	1.0/ 1.0	1.0/ 1.0	1.0/ .56	1.0/ 1.0	.83/ .51	1.0/ .42	.82/ 1.0	1.0/ .50	0.0/ 0.0	.67/ 1.0	1.0/ 1.0	.67/ .67	.50/ .67	1.0/ .33	1.0/ .50	.79	.87/ .59

*The accuracy rate is given first and the criterion accuracy rate is given second:

Accuracy Rate / Criterion Accuracy Rate

Accuracy Rate: Proportion correct of the total recorded

Criterion Accuracy Rate: Proportion of times the criterion instances were recorded correctly.

utilized. Therefore, only codes with six or more examples are considered in the analysis of specific grade levels within a site.

As an illustration of how Table 2 can be used, the results of the first observers listed are discussed. The observers are grouped according to the site or project they observed. Going from left to right, the "What" codes are first shown on the extreme left with the codes which are represented by six or more criterion instances. The next section includes the "What" codes that were represented by fewer than six instances. The "How" codes are shown next, with a similar division.

1. Findings from "What" Codes Occurring Six or More Times

The first observer listed, Observer No. 1 from Site A, had an overall reliability rate of .84 on the "What" codes (see Table 2). Of the nine codes with six or more criterion instances, only two codes registered an "observer accuracy" or "criterion accuracy" rate of less than .75. Looking at code "4," instruction, we see an observed accuracy rate of .62 and a criterion accuracy rate of .89. This means that when Observer 1 recorded the "4" code, it was correct 62 percent of the time. The observer actually recorded a "4" 89 percent of the time; thus, she missed only 11 percent of the examples. However, 38 percent of the time when she recorded "4's" she was incorrect. Therefore, variables using the "4" code in the first grade at Site A should be interpreted with caution.

The other code which the observer's results show to be considered less than adequate was the "5NV" code, nonverbal general action. The accuracy rate of 1.00 shows that when she recorded a "5NV" it was always a "5NV"--she did not confuse it. However, she failed to code 50 percent of the videotape examples of "5NV."

The overall results for Observer No. 1 show that the observation data she gathered can be analyzed with a great deal of confidence. Only the "4" and "5NV" code results have to be analyzed with special caution.

Three of the other first grade observers for this sponsor registered accuracy rates of over .70 on the "4" code. The first grade observer at Site E has an accuracy rate of only .54. If the results of the data collection show that Grade 1, Sites A and E, have means and standard deviations for code "4" that differ widely from the other sites,

it may be explained by the observers' confusion in the use of the "4" code.

A similar situation exists with the data for the four other first grade observers on the "5NV" code. The underestimation of the code by Observer No. 1 at Site A is not common to all first grade observers. Therefore, this should be taken into consideration when the data is analyzed.

2. Findings from "How" Codes Occurring Six or More Times

It can be seen on Table 2 that Observer 1 at Site A was 100 percent accurate when she recorded five of the more frequent "How" codes. The one coding exception is "A," academic. Only 67 percent of the time were her "A" recordings correct. Thirty-three percent of the time they were not "A's." However, she recorded 90 percent of the "A's" actually occurring on the videotape. The extra 33 percent that she recorded are considered intrusions,* and they overestimate the occurrence of this code. Observers at other sites had their own specific difficulties, and their data will have to be analyzed in the same way that Observer 1's has been analyzed.

3. Summary

The usefulness of this method of measuring the accuracy of individual observers lies in its capacity to:

- Differentiate codes according to relatively high or low levels of confidence;
- Assess an individual's coding skill on a specific code and examine observer bias;
- Compare individual observer's scores with other observer's scores at the sponsor's same grade level.

By thus identifying the various sources of error in the observation measures, we can more accurately determine whether specific problems lie in the code itself or with the individual observer and interpret the data accordingly.

*See page 18 for an explanation of intrusion.

E. A STUDY COMPARING INTER-RATER ACCURACY AND VIDEOTAPE SIMULATION ACCURACY

The preceding section has examined the confusability of the observation codes and the ability of observers to code criterion videotapes. Videotaped simulation of classroom events are, admittedly, different from actual classroom events. In an effort to compare the accuracy of observer ratings on the simulations and inter-rater accuracy in classrooms, a small study was conducted in one location. This section compares the results obtained from both studies of accuracy for two observers.

1. Paired Observers

The first method, the paired observation, is the most commonly used method of assessing interaction analysis instruments. The procedure followed is to have the two observers situated in the same classroom, coding exactly the same situation simultaneously. The recorded codes are then evaluated in terms of percent agreement between the two observers. Since the speed of the two observers is not expected to be consistent, the ratio of the number of codes recorded by the observer is compared to the ratio of the number of codes recorded by the trainer.

It must be pointed out that this paired observation procedure has some serious limitations. First, two extra people in the classroom are more obtrusive than one. Second, it is almost impossible to assure that the two observers are focusing on exactly the same action. Due to limited space, the two observers may not have the same angle of observation; thus, what they see and hear may be somewhat different and yet each observer could be collecting a correct and adequate sample of the behavior which is occurring. A third problem is that even if the marginal frequency counts of a code by two observers are numerically similar, we cannot be certain that the two observers have recorded specific incidents exactly the same. Similar ratios could occur by chance. Lastly, it happens that during the classroom observations certain interactions or codes do not occur, or occur at such a minimal rate, that reliability cannot be computed. There is no way to be certain that all codes will be assessed within a given time period.

In the study, data from the 16 five-minute observations were

examined, using three "Who" codes, twelve "What" codes, and thirteen "How" codes. To assess the coding accuracy of the two observers, the proportion of frames that contained a particular code was recorded for each trainer and trainee. From the proportions, the following equation was computed for each code (p is for the trainer and q is for a given observer on a given code):

$$\text{The percent agreement} = 100 \times \frac{\min (q,p)}{\max (q,p)}^{*}$$

Tables 3 and 4 show the overall percentage reliability of the codes separately in terms of their ratio of frequency. It must be noted that accuracy for low frequency variables is difficult to interpret because if one observer records an event four times and the other only two times and they observe an equal number of frames, the agreement is only 50 percent, even though the actual difference is only two occurrences. Higher frequency variables can tolerate a difference of two occurrences and still show a high percentage of agreement. The data for each observer is presented separately in Tables 3 and 4. Since there are 16 paired observations, it is possible to have as many as 1,216 frames of interaction. Therefore, we have separated the data into three categories: least frequent, moderately frequent, and most frequent. Table 5 is included to further clarify the results of the paired observations. It includes the frequency scores of the SRI trainer as well as the ratios of occurrence and percent agreement scores for both observers over all codes.

The results show that both of the observers were very reliable on the "Who" codes. The "What" codes were also recorded very reliably, with only two exceptions. Observer 1 recorded less than half as many "8" codes (praise) as the criterion observer, and Observer 2 missed nearly 80 percent of the occurrences of code "6" (task related statement). Significantly, however, both of these codes occurred with low frequency.

The results on the "How" codes were much lower. Observer 1 was quite reliable on the "NV" (nonverbal), "G" (guide to alternative), "A" (academic), and "B" (behavior) codes. She was below the 50 percent

*When p = 0 and q = 0, the percent agreement is assigned a value of 100.

Table 3

PERCENT AGREEMENT BETWEEN TRAINER/OBSERVER 1

WHO CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100			Adult, Child	2
81-90	Machine			1
71-80				
61-70				
51-60				
41-50				
			TOTAL	3

WHAT CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100	10		4	2
81-90		1Q, 6	1, 2, 12	5
71-80	11			1
61-70		5, 9		2
51-60	7			1
41-50	8			1
			TOTAL	13

HOW CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100	U, G, DP		A	4
81-90				
71-80			NV	1
61-70		B		1
51-60				
41-50	X			1
31-40	T			1
21-30				
11-20	Q	H		2
0-10	N, O, W			3
			TOTAL	13

Table 4

PERCENT AGREEMENT BETWEEN TRAINER/OBSERVER 2

WHO CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100			Adult, Child	2
81-90				
71-80				
61-70				
51-60				
41-50				
			TOTAL	2

WHAT CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100	8		5	2
81-90			3	1
71-80		4, 7	10	3
61-70	10		12	2
51-60		9	1	2
41-50				
31-40				
21-30				
11-20	6			1
			TOTAL	11

HOW CODES

Percent Agreement	Least Frequent (0-60)	Moderately Frequent (61-175)	Most Frequent (176-1,216)	Total No. of Codes
91-100	DP		A	2
81-90				
71-80			NV	1
61-70				
51-60	X			1
41-50				
31-40	T			1
21-30	H, U, Q, G			4
11-20		B		1
0-10	N, O, W			3
			TOTAL	13

Table 5

PAIRED OBSERVATION RESULTS

		Observer 2				Observer 1			
		Percent Agreement	Observer Ratio	Trainer Ratio	Trainer Score	Percent Agreement	Observer Ratio	Trainer Ratio	Trainer Score
WHO	Adult	94	.544	.578	638	97	.482	.467	502
	Child	98	.431	.423	467	96	.488	.506	544
	Machine	--	.026	.000	0*	82	.023	.028	30*
	TOTAL FREQUENCY		990	1,105	1,105		1,160	1,076	1,076
WHAT	1	60	.174	.104	115	88	.171	.150	161
	1Q	73	.106	.145	160	90	.103	.093	100
	2	--*	.001	.000	0*	--*	.000	.001	1*
	3	89	.288	.257	283	88	.212	.242	260
	4	77	.091	.070	77	99	.150	.151	162
	5	92	.091	.099	109	65	.086	.056	60
	6	17*	.006	.035	38*	81	.069	.056	60
	7	75*	.039	.052	57*	56*	.020	.036	38*
	8	100*	.022	.022	24*	44*	.007	.016	17*
	9	55	.038	.069	76	65	.042	.065	69
	10	65*	.031	.020	22*	93*	.013	.014	15*
	11	--*	.020	.000	0*	75*	.003	.004	4*
	12	70	.087	.124	137	83	.121	.100	107
	NV	71	.144	.203	224	73	.239	.174	187
	X	51*	.021	.041	45*	43	.053	.023	23*
HOW	H	22*	.004	.018	20*	11	.006	.057	61
	U	25*	.016	.004	4*	100*	.000	.000	0*
	N	0*	.000	.009	10*	8*	.001	.012	13*
	T	33*	.003	.009	9*	33*	.003	.009	9*
	Q	23*	.005	.022	24*	11*	.001	.009	9*
	G	29*	.008	.028	31*	100*	.033	.033	35*
	P	--*	.002	.000	0*	--*	.000	.001	1*
	O	--*	.000	.001	1*	0*	.000	.016	17*
	W	--*	.000	.002	2*	0*	.000	.007	7*
	DP	100*	.000	.000	0*	100*	.000	.000	0*
	A	93	.670	.622	687	94	.760	.713	767
	B	16*	.008	.051	56*	62*	.018	.029	31*

*Fewer than 60 criterion instances.

Ratio = occurrence of a specific code/total number of frames recorded

agreement rate for the "X" (movement), "H" (happy), "N" (negative), and "O" (object) codes. The remaining codes occurred less than ten times and, therefore, no accuracy rate could be arrived at.

Observer 2's rate of accuracy was similar on the "How" codes. She was reliable on the "NV," "DP," and "A" codes and below the 50 percent level on the "Q," "G," and "B" codes. Eight of the "How" codes occurred only ten or fewer times, thus generalizations regarding these codes would be made with caution.

2. Videotaped Skits (Simulations)

The second phase of the reliability study was based on videotaped skits. The procedure followed is to have the observers code interactions seen on a videotape and compare that record with predetermined criteria. The tape has stops or pauses between each interaction to insure that each observer knows which interaction to code.

The results are then compiled for each observer, and they reveal both (1) which occurrences were not recorded and (2) which code was erroneously recorded in its place. The procedure allows us to identify the problem codes for each specific observer.

In the figures that follow, two values are shown in each cell. For those cells that fall on the main diagonal, the upper value shows the percent of times the total number of codes recorded was correct. The lower value shows the percent of times the code actually occurred and was recorded correctly by the observer.

For cells that do not fall on the diagonal, the two values indicate proportions of error rather than of accuracy. The upper value shows the percent of times a specific code (as shown by row indicator) was recorded instead of a specific criterion code (indicated by the column) to the total number of recordings of that code. The lower value indicates the percent of times that the specific code (indicated by the row) was recorded when a given criterion code was called for (shown by the column).

Figures 10, 11, 12, and 13 are matrices showing the percent of accuracy and the percent of the total codes recorded for the two observers. Computations are for the "What" and the "How" codes. The total number of criterion instances of each code is shown at the bottom of each column.

Figure 10

PROPORTIONAL DISTRIBUTION OF "WHAT" CODES FOR OBSERVER 1 FOR BOTH RECORDED AND CRITERION ACCURACY

Criterion Examples (Lower Numbers in Cells)

"What" Codes (Upper Numbers in Cells)	1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	11	12	Total Entries	Total Correct Entries
1	.67 .50							.33 1.00								3	2
1Q	.08 .50	.72 .86	.20 1.00													25	18
2			0.00 0.00													0	0
3				.96 .89		.04 .10										24	23
3NV					1.00 1.00											4	4
4		.13 .05				.88 .70										8	7
4NV							.33 .33		.67 .33							6	2
5				.13 .04				0.00 0.00		.88 .54						8	0
5NV							.36 .67		.64 .58							11	7
6		.20 .05								.80 .31						5	4
7				.18 .08							.82 .90					11	9
8												1.00 1.00				2	2
9										.10 .08			.90 1.00			10	9
11														1.00 1.00		2	2
12		.03 .05				.07 .20			.03 .08	.03 .08	.03 .10				.79 1.00	29	23
Total Number of Examples on Criterion Tapes	4	21	5	26	4	10	6	1	12	13	10	2	9	2	23 →	148	112 (76% accuracy rate)

Figure 11

PROPORTIONAL DISTRIBUTION OF "WHAT" CODES FOR OBSERVER 2 FOR BOTH RECORDED AND CRITERION ACCURACY

Criterion Examples (Lower Numbers in Cells)

"What" Codes (Upper Numbers in Cells)	1	1Q	2	3	3NV	4	4NV	5	5NV	6	7	8	9	11	12	Total Entries	Total Correct Entries
1	.22 .50	.44 .27	.11 .20										.22 .22			9	2
1Q	.08 .25	.85 .73	.08 .20													13	11
2			1.00 .40													2	2
3				.89 .82	.04 .33					.08 .15						26	23
3NV					1.00 .67											2	2
4				.20 .04		.80 .44										5	4
4NV							.80 .40			.20 .08						5	4
5	.14 .25						.43 .30	0.00 0.00		.29 .15			.14 .11			7	0
5NV							.25 .30		.75 1.00							12	9
6			.07 .20	.14 .07						.57 .62	.07 .10				.14 .09	14	8
7											.75 .60	.25 1.00				8	6
8												0.00 0.00				0	0
9													1.00 .67			6	6
11														1.00 1.00		3	3
12				.07 .07		.17 .56					.10 .30				.67 .91	30	20
																	100
Total number of Examples on Criterion Tapes	4	15	5	28	3	9	10	0	9	13	10	2	9	3	22	142	(70% accuracy rate)

Figure 12

PROPORTIONAL DISTRIBUTION OF "HOW" CODES FOR OBSERVER 1 FOR BOTH RECORDED AND CRITERION ACCURACY

"How" Codes (Upper Numbers in Cells)	Criterion Examples (Lower Numbers in Cells) NV	X	A	B	DP	W	O	P	G	Q	T	N	U	H	Proportion of Intrusions		Total Entries	Total Correct Entries
NV	.96 .89														.04 .05		25	24
X		.80 .80													.20 .05		5	4
A			.94 .92												.06 .10		36	34
B				.90 .75											.10 .05		10	9
DP					.86 1.00										.14 .05		7	6
W						.50 .50									.50 .05		2	1
O							.36 .50								.64 .33		11	4
P								0.00 0.00							1.00 .05		1	0
G									.14 1.00	.14 .50		.14 .50			.57 .19		7	1
Q										.50 .50					.50 .05		2	1
T											1.00 .33						1	1
N												0.00 0.00			1.00 .05		1	0
U													1.00 .25				1	1
H														1.00 .50			1	1
Proportion of Omissions	.14 .11	.05 .20	.14 .08	.14 .25		.05 .50	.18 .50				.09 .67	.05 .50	.14 .75	.05 .50	0.00 0.00		110	87 (78% accuracy rate)
Total Number of Examples on Criterion Tapes	27	5	37	12	6	2	8	0	1	2	3	2	4	2 →		111		

Figure 13

PROPORTIONAL DISTRIBUTION OF "HOW" CODES FOR OBSERVER 2 FOR BOTH RECORDED AND CRITERION ACCURACY

Criterion Examples (Lower Numbers in Cells)

"How" Codes	NV	X	A	B	DP	W	O	P	G	Q	T	N	U	H	Proportion of Intrusions		Total Entries	Total Correct Entries
NV	.94 .59														.06 .04		17	16
X		1.00 .83															5	5
A			.63 .77	.07 .23											.30 .52		43	27
B				0.00 0.00													0	0
DP					0.00 0.00												0	0
W						0.00 0.00											0	0
O							1.00 .86										6	6
P								0.00 0.00							1.00 .40		10	0
G									0.00 0.00			.50 1.00			.50 .04		2	0
Q										1.00 .50							1	1
T											1.00 1.00						2	2
N												0.00 0.00					0	0
U													1.00 .50				2	2
H														1.00 .50			1	1
Proportion of Omissions	.25 .41	.02 .17	.18 .23	.23 .77	.16 1.00	.05 1.00	.02 .14			.02 .50			.05 .50	.02 .50	0.00 0.00		89	60 (56% accuracy rate)
Total Number of Examples on Criterion Tapes	27	6	35	13	7	2	7	0	0	2	2	1	4	2 →		108		

Codes (Upper Numbers in Cells)

The total number that the observer recorded is given at the end of each row.

Those codes that occurred fewer than seven times are listed in the matrices but will not be discussed in the body of this text. A decision was made that, in these cases, the confidence level with which we might make predictions as to the reliability of an observer would be so low as to render it unacceptable. Therefore, only the codes which were tested by seven or more criterion examples will be considered in this analysis.

As shown in Figure 10, the "What" matrix for Observer 1 indicates that, of the eight codes that included seven or more criterion instances, only the "6" code (task-related statement) and the "5NV" code had a criterion accuracy lower than .70. In the "6" code, 80 percent of the recorded "6" codes were correct, but 69 percent of the criterion codes were missed. Moving up the "6" column we can see that 54 percent of the criterion "6" codes were incorrectly coded as "5." The problem with the "5NV" code is somewhat different. In this case the problem is that both the criterion rate and the observer correctness were low. It appears that on the simulations Observer 1 had difficulty distinguishing the "5NV" code from the "4NV" (self learning or instruction) code, since she often codes the criterion "4NV" instances as "5NV" and also the "5NV" criterion as "4NV."

Over all "What" codes, Observer 1 is reasonably accurate with a criterion rate of .76 which is average for all 63 observers examined by the videotapes on the "What" codes.*

Observer 2 had a reasonable overall criterion accuracy rate (.70) also, but she had coding problems with several codes (see Figure 11). She did not record the "4" code (instruction) 66 percent of the time. The "4NV," "6," "7," and "9" codes were also coded less frequently than required. She used the "12" code (observing) eight more times than required. They were confused with codes "3," "4," and "7."

The "How" code accuracy for Observer 1 was also acceptable (see Figure 12). Her overall criterion accuracy rate was .78. This figure indicates that of the 111 criterion "How" codes presented, she recorded them correctly 87 times (see lower right hand corner of Figure 12).

*This figure is computed by dividing the total number of correct entries by the exact number of videotaped criterion examples.

The only "How" code that Observer 1 recorded with less than 70 percent accuracy was the "O" code (objects); 50 percent were missed and 64 percent were recorded when not indicated. The other codes that fell below a .70 rate of accuracy were codes that included fewer than seven criterion instances.

Observer 2 had a more difficult time recording the "How" codes from the tapes. In Figure 13, her overall accuracy rate is shown as only .56. On individual codes, the "O" was very reliable (1.00/.86), but the "NV" was not used 41 percent of the time required. The "A" (academic) was coded when not called for sixteen times as well as omitted eight times when it should have been coded. "B" (behavior) and "DP" (dramatic play) were ignored completely.

3. Conclusions

Two distinct procedures, the videotaped skits and the paired observations, were used to assess the accuracy of two observers. The results indicate average reliability for both observers on the "What" code category. For the "How" category, Observer 1 is above average, but Observer 2 is below the average of the other 62 observers.

Specifically, Observer 1 was acceptably accurate on the more frequently used individual codes. Many of her codes, such as the "1Q," "3," "4," "7," "9," "12," "NV," "A," and "B" were shown to be very reliable on both procedures. Only the "O" code (use of objects) was shown to be unreliable on both procedures.

The results were equally good for Observer 2 on the "What" codes with only the "6" code (task related comment) being shown unreliable on both procedures. The "How" code "B" (behavior) was also recorded poorly in both procedures. In the case of the videotape codings she missed the 13 examples of the "B" code and underestimated it in the paired observations. On the "A" code (academic), Observer 2 was 93 percent accurate on the paired observations but had a .63/.77 reliability on the videotapes. Other "How" codes such as "movement" and "object" are acceptably accurate on the videotapes while "nonverbal" is acceptably accurate on the paired observations. "Guide" and "question," which were underestimated in the inter-rater analysis, have too few examples on the videotape to be discussed in terms of reliability.

While simulated videotaped events are limited in their scope

and differ from the classroom situation, they do offer a standard stimulus to examine each observer's ability to code specified events and to identify observer bias. There is still some confounding in the source of "system error"; however, the variation introduced by a second observer is eliminated. While the two systems of examining observer accuracy do yield some different information, it is not contradictory, and the videotape system is by far easier to interpret.