

Effects of Definition Characteristics on the Cross-Laboratory  
Generalizability of Direct Observation Data

Bernard D. Fabry  
Boys Town

Robert P. Hawkins  
Sharon L. Foster  
David Brunts  
West Virginia University

Presented at the Eighth Annual Convention of the Association for Behavior Analysis,  
Milwaukee, May, 1982.

Effects of Definition Characteristics on the Cross-Laboratory  
Generalizability of Direct Observation Data

Bernard D. Fabry  
Boys Town

Robert P. Hawkins  
Sharon L. Foster  
David Brunts  
West Virginia University

Using observers to collect behavioral data has long been recognized as a potential source of measurement error (Arrington, 1943; Kent & Foster, 1977). One aspect of the problem which has received little attention is the extent to which error can be attributed to differences in laboratory settings (Foster & Cone, 1980). Two studies (DeMaster, Reid & Twentyman, 1977; Wildman, Erickson & Kent, 1975) have provided suggestive data. In those studies interobserver agreement was found to be higher within groups of observers trained together than across groups. The results suggest that when a behavioral coding system is developed and used in one laboratory, observers in other laboratories may not learn to use the system in a similar manner. One purpose of the present study was to evaluate the extent to which data obtained from a laboratory in which a behavioral coding system was developed would be generalizable across other laboratories.

A second purpose of the study was to determine the extent to which the definition quality of observational categories affects observer error. A few studies have obtained results suggesting that carefully constructed, objective definitions improved observer agreement (Frame, 1979; Hawkins & Dobes, 1977).

### Method

To assess the effects of different laboratory settings on observers, undergraduate college students were recruited and randomly assigned to one of three training groups. Training for one group of five observers was designed to be analogous to training obtained in a "parent" laboratory which had originated a coding system. The trainer had developed a set of observational categories and was experienced in training observers. Training for the other two groups of five observers each was designed to be analogous to training at separate laboratories, or locales, that had chosen to adopt the behavioral categories and coding system. One of those groups, the "veteran" laboratory, was trained by an experienced trainer. Observers in the third group, the "novice" laboratory, trained themselves.

To assess the effects of definition quality the trainer for the parent laboratory had written five definitions according to a set of rules derived from research on concept learning (Markle & Tiemann, 1972). The rules specified that each definition include (1) critical relevant variables, (2) important irrelevant variables, and (3) a rational set of exemplars and nonexemplars. These five definitions were referred to as maximal definitions. Another five definitions were written to be equally clear and objective (Hawkins & Dobes, 1977), but not as complete. Those definitions, referred to as minimal definitions, included only a description of critical relevant attributes. They were similar to definitions typically found in the literature.

The groups of observers watched and coded the same sequence of videotapes during training and used the same set of definitions so that their protocols, or coding sheets, could be compared. The effects of laboratory setting were assessed by comparing interobserver agreement scores obtained both across observers within laboratories as well as across observers in different laboratories. The effects of definition quality were assessed by comparing agreement scores obtained with the maximal definitions versus agreement scores obtained with the minimal definitions.

To assess changes in observer performance as training progressed, the first three training sessions were used in all analyses to represent observer performance at the start of training. The last three training sessions were used to represent performance at the end of training. Finally, a three-session post training phase was included. During that phase all observers worked independently of the trainers and of each other as would occur during routine data collection.

## Results

Table 1 compares the two sets of definitions on five interobserver agreement coefficients across the three phases of training. Median agreement scores were computed for all pairwise comparisons within each laboratory and definition type. As can be seen, agreement was generally low at the start of training for the minimally defined categories, rose by the end of training, and then decreased somewhat in the post training phase. Similar results were obtained for all three laboratories.

With respect to the categories which were defined as concepts (maximal definitions), agreement at the start of training was somewhat higher than for the minimal definitions. By the end of training, agreement was not substantially different, though. During the post training phase, agreement was somewhat higher for maximal definitions than for minimal definitions on Kappa and weighted agreement (Harris & Lahey, 1978) for two of the groups.

Table 2 shows between laboratory agreement coefficients for minimal and maximal definitions. Across all three phases of training, agreement was higher for maximal definitions than for minimal definitions. Other analyses on interval scores and session totals, including generalizability analyses as developed by Cronbach, Gleser, Nanda and Rajaratnam (1972), support the results presented in the present paper.

## Discussion

The results indicated that cross-laboratory comparability of data cannot be assumed. Observers within laboratories agreed well among themselves but did not agree as well with observers working in other laboratories. However, defining behavioral categories according to rules for defining concepts may ameliorate that effect. The observers agreed more closely across laboratories on the maximal definitions than they did on the minimal definitions. In addition, agreement on the maximal definitions was higher at the start of training and post-training than on the minimal definitions both within and across laboratories.

## References

Arrington, R. E. Time sampling in studies of social behavior: A critical review of techniques and results with research suggestions. Psychological Bulletin, 1943, 40, 81-124.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measures. New York: Wiley, 1972.

DeMaster, B., Reid, J., & Twentyman, C. The effects of different amounts of feedback on observers' reliability. Behavior Therapy, 1977, 8, 317-329.

Foster, S. L., & Cone, J. D. Current issues in direct observation. Behavioral Assessment, 1980, 2, 313-338.

Frame, R. E. Interobserver agreement as a function of the number of behaviors recorded simultaneously. Psychological Record, 1979, 29, 287-296.

Harris, F. C., & Lahey, B. B. A method for combining occurrence and non-occurrence interobserver agreement scores. Journal of Applied Behavior Analysis, 1978, 11, 523-527.

Hawkins, R. P., & Dobes, R. W. Behavioral definitions in applied behavior analysis: Explicit or implicit. In B. C. Etzel, J. M. LeBlanc and D. M. Baer (Eds.), New developments in behavioral research: Theory, methods, and applications. I honor of Sidney W. Bijou. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977.

Kent, R. N., & Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. Calhoun, and H. E. Adams (Eds.), Handbook of behavioral assessment. New York: Wiley, 1977.

Markle, S. M., & Tiemann, P. W. (Some principles of instructional design at higher cognitive levels.) In K. Weltner and B. Rolett (Eds.) Fortschritte und ergebnisse der unterrichtstechnologie. Berlin: Ehrenwirth-Verlag, 1972. (In R. Ulrich, T. Stachnik, and J. Mabry (Eds.) Control of human behavior (Vol. 3). Glenview, Ill.: Scott, Foresman, 1974.)

Wildman, B. G., Erikson, M. T., & Kent, R. N. The effect of two training procedures on observer agreement and variability of behavior ratings. Child Development, 1975, 46, 520-524.

Table 1  
Median interobserver agreement coefficients within laboratories

Coefficient	Start of Training		End of Training		Post Training	
	Minimal Definition	Maximal Definition	Minimal Definition	Maximal Definition	Minimal Definition	Maximal Definition
<u>Parent Lab</u>						
I-by-I	.877	.899	.877	.938	.857	.856
Kappa	.289	.509	.429	.379	.346	.459
Occurrence	.267	.462	.357	.308	.375	.479
Nonoccurrence	.870	.878	.851	.915	.819	.814
Weighted	.294	.528	.444	.374	.392	.510
<u>Veteran Lab</u>						
I-by-I	.949	.894	.971	.974	.977	.937
Kappa	0	.292	.441	.340	.498	.467
Occurrence	0	.250	.375	.253	.499	.334
Nonoccurrence	.975	.889	.978	.979	.979	.936
Weighted	.070	.344	.439	.336	.515	.392
<u>Novice Lab</u>						
I-by-I	.931	.901	.987	.978	.991	.883
Kappa	.194	.334	.396	.509	.001	.396
Occurrence	.200	.309	.419	.429	.001	.376
Nonoccurrence	.935	.892	.979	.979	.999	.875
Weighted	.266	.374	.450	.508	.030	.438
<u>Mean</u>						
I-by-I	.919	.898	.945	.963	.942	.892
Kappa	.161	.378	.422	.409	.282	.441
Occurrence	.156	.340	.384	.330	.292	.396
Nonoccurrence	.921	.886	.936	.958	.932	.875
Weighted	.210	.415	.444	.406	.312	.447

Table 2  
Median interobserver agreement coefficients between laboratories

Coefficient	<u>Start of Training</u>		<u>End of Training</u>		<u>Post Training</u>	
	Minimal Definition	Maximal Definition	Minimal Definition	Maximal Definition	Minimal Definition	Maximal Definition
<u>Parent Lab</u>						
I-by-I	.895	.861	.851	.926	.872	.817
Kappa	.049	.308	.060	.032	.099	.269
Occurrence	.077	.313	.071	.063	.094	.231
Nonoccurrence	.893	.846	.833	.917	.851	.804
Weighted	.173	.392	.173	.159	.189	.323
<u>Veteran Lab</u>						
I-by-I	.937	.881	.969	.968	.975	.877
Kappa	0	.279	0	.053	0	.175
Occurrence	.046	.261	0	.100	0	.176
Nonoccurrence	.937	.875	.959	.959	.979	.872
Weighted	.125	.369	.047	.178	.029	.249
<u>Novice Lab</u>						
I-by-I	.888	.876	.819	.945	.869	.840
Kappa	.158	.326	0	.379	0	.283
Occurrence	.166	.333	0	.333	0	.286
Nonoccurrence	.875	.853	.792	.937	.851	.809
Weighted	.245	.392	.103	.414	.099	.366
<u>Mean</u>						
I-by-I	.907	.873	.880	.946	.905	.845
Kappa	.069	.304	.020	.155	.033	.242
Occurrence	.096	.302	.024	.165	.031	.231
Nonoccurrence	.902	.858	.861	.938	.894	.828
Weighted	.181	.384	.108	.250	.106	.313

This research was approved by a human subjects committee at West Virginia University, Morgantown, West Virginia

## Appendix A

## Rules for Defining Behavioral Categories as Concepts

Each definition should include:

1. A numbered list of critical attributes. A critical attribute is one aspect or component of the target response class that is necessary (but not necessarily sufficient) for a response to be considered an instance of the target response class. Critical attributes are conjunctive attributes; they can be conjoined by “and.”
2. A numbered list of important variable attributes. An important variable attribute is an aspect or feature of a response that is not a necessary part of the target response class. Variable attributes are disjunctive attributes that would be conjoined by “or.”
3. A rational set of exemplar/nonexemplars pairs plus rationales for each pair. A rational set consists of a broad range of exemplars and nonexemplars carefully selected to eliminate any wrong hypotheses that an observer may entertain. More specifically, a rational set consists of:
  - a. Enough exemplars such that each variable attribute is illustrated, and no variable attribute is present in every exemplar;
  - b. The variable attributes illustrated by the exemplars are as different as possible across the exemplars, and there is no consistent pattern of pairings of the variable attributes across the exemplars (as far as possible);
  - c. One nonexemplar is paired with each exemplar such that the nonexemplar includes variable attributes that match as closely as possible the variable attributes of the exemplar; and
  - d. Each nonexemplar omits one critical attribute not omitted by any other nonexemplars.
4. The rationale for each exemplar/nonexemplars pair indicates the variable attributes included in both the exemplar and nonexemplars plus the critical attribute omitted from the nonexemplars.

Appendix B

An example of a minimally defined category

Playing – symbol = P

Components

Child uses his hands to play with his own or community property, so that such behavior is incompatible (or would be incompatible) with learning.

## Appendix C

## An example of a maximally defined category

Time Off Task – symbol = X

Critical Components

1. Child does not look at (have his head oriented in the direction of) his work, and
2. does not use an appropriate implement (if required for the work)
3. for the entire “observe” interval, and
4. his hand is not raised,
5. the work is assigned and
6. he has not been told he may cease working.

Variable Components

7. The assigned work (7a) may or (7b) may not be finished.
8. The teacher (8a) may be presenting a lesson, or (8b) another child may be presenting information to the class, or (8c) no one is presenting anything.

-----

Examples

Include:

Child has a reading book in front of him which he was told to read. But he watches the teacher for the entire interval while she presents information on the blackboard.

Exclude:

Child has a reading book in front of him which he has not been looking at but was assigned to read. During the interval the teacher tells him to pay attention to her (stop reading is implied) while she presents information on the blackboard.

For both examples the child was not doing reading (1) which had been assigned (5) for an entire interval (3), and did not raise his hand (4); no implement was required for the work (2). Only the include example indicates that he was not told to stop working (6). The facts that the work was not finished (7b) or that the teacher was presenting a lesson (8a) are not important.

-----

Include:

Child has been assigned a cutting and pasting task. He does not cut and paste and just stares at his work for an entire interval.

Exclude:

Child has been assigned a cutting and pasting task. He does not look at the work, but cuts and pastes only momentarily.

For both examples the child was looking at work (1) which has been assigned (5); his hand was not raised (4), and the teacher did not tell him to stop working (6). The fact that he was looking at his work would exclude both of these examples except that he did not use required implements (2) for an entire interval in the include example. The facts that the work was not finished (7b) or that no one was presenting anything (8c) are not important.

-----

Include:

Child has been assigned a writing task in his workbook but stares at another child who is answering a question for the teacher.

Exclude:

Child has been assigned a writing task in his workbook, but raises his hand to give a different answer to a question being answered by another child.

Both examples describe a child not looking at his work (1) and not using a required implement (2) for an entire interval (3); the work had been assigned (5) and he had not been told he could stop working (6). But only the include example indicates that his hand had not been raised (4). The facts that the work was not finished (7b) or that another child was presenting information (8b) is not important.

-----

Include:

Child has finished the work on one page of a workbook, turned the page, and is now watching another child working.

Exclude:

Child had previously been told to turn the page of his workbook when he finished it (implicit direction to stop working when page is finished). The child did that during an earlier interval and had not been given a new assignment. He is watching another child work.

For both examples the child does not look at his work (1) or use a required implement (2) for an entire interval (3), and his hand is not raised (4). But only in the left example has work been

assigned (5) and the child not told to stop working (6). The facts that the work is finished (7a) or that no one is presenting anything (8c) are irrelevant.

-----