WILDMAN, BETH GOLDSTEIN. A Comparison of Two Training Procedures for Maintaining Inter-Rater Reliability. (1974) Directed by: Dr. Marilyn T. Erickson. Pp. 34.

Much behavior modification research has relied on the use of human observers to collect data. However, instrumental errors have been found to be associated with the collection of observational data. The present study compared two types of training procedures in which the consistency of the standard, to which the raters were trained to conform, was varied.

Sixteen college undergraduates (Ss) without prior experience in observing classroom behavior were trained in the observation of nursery-school children's behavior.

All Ss viewed the same 40-minute videotapes of nursery-school children. Seven 60-minute training sessions were conducted using the O'Leary disruptive behavior code. Four pairs of observers in Group I were trained by one graduate student trainer whose ratings were accepted as the standard. Four pairs of observers in Group II were trained by themselves, thus establishing their own standard.

Following training, six videotapes were rated by both groups. Videotapes were divided into four 10-minute blocks to permit the collection of overtly and covertly assessed reliabilities for both within-pair and between-pair combinations.

Overtly assessed reliabilities were found to be significantly higher than covertly assessed reliabilities

($p < .05$). The reliabilities within observer pairs was found to be significantly higher than reliabilities between observer pairs ($p < .05$). The change of reliabilities over time was found to be significant, with the reliabilities for the last two sessions higher than reliabilities for the first two sessions ($p < .01$).

Group I rated significantly more behaviors than Group II ($p < .01$), suggesting that the groups were applying the disruptive behavior code differently.

The results of the present study corroborate the existence of instrumental errors associated with the use of human observers. In addition, the results indicate that the type of training procedure used may affect the way the code is applied. Thus, unsystematic methods for training observers may lead to different observers applying behavior codes differently. In addition, observer-pairs trained by the same procedure may not be reliable with one another.

A COMPARISON OF TWO TRAINING PROCEDURES FOR MAINTAINING INTER-RATER RELIABILITY

by

Beth Goldstein Wildman

A Thesis Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Greensboro
1974

Approved by

Marilyn T. Erickson
Thesis Adviser

APPROVAL PAGE

This thesis has been approved by the following committee of the Faculty of the Graduate School at The University of North Carolina at Greensboro.

Thesis Adviser Marilyn T. Erickson

Committee Members A. R. Sedesquist

Robert M. Pratt

February 20 1974
Date of Acceptance by Committee

ACKNOWLEDGMENTS

The author gratefully acknowledges her chairwoman, Dr. Marilyn T. Erickson, for her guidance and support through all phases of this research. Thanks are due to Dr. David R. Soderquist and Dr. Robin W. Pratt for their helpful suggestions in the preparation of this thesis.

The author appreciates the time her subjects: Glenda Akin, Carol Blaine, Sandra Caudle, Shirley England, Janina Garner, Delores Gilbert, Charlotte Goins, Hope Haywood, Theresa Holloway, Gail McCormick, Jean Merwin, Bette Rausch, Patricia Shaw, Keritha Shore, Peggy Townsend, and Sylvia Wright, devoted to her research.

The author wishes to thank the teachers and children of Carter Child Care Center for their patience and toleration of the inconveniences of being videotaped.

The author gratefully acknowledges Dr. William Powers for his help with the statistical analysis and computer programming.

The suggestions of Dr. Ronald Kent on the design and content of this research were indispensable.

The assistance of Hal Wildman in reviewing all drafts of this thesis and in the statistical analysis is also appreciated.

TABLE OF CONTENTS

LIST OF TABLES

Introduction

Much of the current research in the area of behavior modification has relied on the use of human observers, typically college undergraduates, to collect the necessary data. The data collected by these observers are frequently assumed to be as objective as electronic recording devices placed in the classroom to record the behaviors of the children. The data collected by these observers, however, may be considerably less consistent than would be expected from precise electronic recording devices.

Minimization of observational errors and the necessity for an objective means of recording behaviors which occur in the natural environment has long been recognized as a necessity for implementing observational research (Olson, 1929; Loomis, 1931; Arrington, 1932). Early studies in the area of child development focused on the development of observation techniques, such as time-sampling (Olson, 1929). Olson (1929), Loomis (1931), and Arrington (1932) were pioneers in the use of simultaneous observations from at least two observers. These early studies initiated the current observational methodology. Arrington (1932) identified several of the major factors affecting inter-observer reliabilities: lack of clear definitions of behaviors; differences in the perspective of the

observers, such as different angles of observation or different distances from the child; observer bias; and problems in the exact timing of intervals.

Sophistication of observational techniques, as well as advancements in technology, such as videotapes, small portable stopwatches, and precise operational definitions, have eliminated many of the difficulties reported by researchers in the 1920's and 1930's. Investigators in the 1960's, however, have employed observational techniques which rely on human observers without consideration of the remaining inadequacies inherent in these data.

Observers are trained to recognize the occurrence or non-occurrence of certain operationally-defined behaviors. After practicing observation and recording of these behaviors until they have obtained some predetermined level of agreement with other observers, they are sent into a field experimental situation to collect data assessing the effectiveness of various experimental manipulations. In order for confidence to be placed in the data, the consistency of the recording of the observed behaviors over time and between observers must be demonstrated, in order to insure that the data do not reflect the idiosyncratic judgements of the observers.

For this reason, reliability (percent of agreement, such as number of agreements divided by number of agreements plus number of disagreements) is periodically assessed

throughout the course of an experiment by having at least two observers monitoring the behaviors of the subject(s) simultaneously. The reliability reflects the amount of agreement between observers.

Renewed interest in the validity and reliability of observational data has occurred in the 1970's (Kass & O'Leary, 1970; Skinrud, 1973; Jones, 1973). Recent studies (Reid, 1970; Romanczyk, Kent, Diament, & O'Leary, 1973) and reviews (Johnson & Bolstad, 1973; O'Leary & Kent, 1973) have yielded empirical as well as ad hoc evidence that observational data are not necessarily objective and may, in fact, suffer from observer errors.

Reid (1970) empirically demonstrated that the assessment of reliability is a reactive process. The results of his study indicated that reliabilities obtained from observers who are informed that reliability is being assessed are considerably higher than reliabilities obtained from observers who are unaware that reliability is being measured.

Romanczyk et al. (1973) expanded upon the difficulties inherent in obtaining "true" measures of reliability. The Romanczyk et al. study presented strong evidence indicating that not only does knowledge of reliability assessment affect obtained reliabilities, but knowledge about who assessed the reliability also increased the obtained reliabilities. Observers appeared to invoke the

idiosyncratic definitions of certain of the behavioral codes in order to conform to the rating standards of their reliability assessor. Reliabilities obtained with a known assessor were much higher than those attained when reliability assessment was obtained with an unknown assessor.

The reliabilities reported in most studies which assess reliabilities with a known assessor at known times are thus not random samples of reliabilities. In fact, the reliabilities reported from occasional assessments are probably considerably higher than they are during periods of non-assessment.

Consistent observation standards are particularly necessary when multiple raters are employed. Studies that employ different groups of raters to observe the same subject, i.e., different raters for morning and afternoon or for different days of the week, would need groups of observers that are reliable with each other. Between-groups experimental designs present particular problems in that different groups of observers are usually assigned to a particular classroom, school, hospital, or home. In these situations, it is difficult to assess the reliability that one set of observers has with another.

O'Leary and Kent (1973) presented evidence suggesting that, although members of a group may obtain high within-group reliabilities, there may be a significant difference in the ratings between groups. O'Leary and Kent reported:

> It seems clear that the magnitude of the differences is sufficient to distort treatment effects, had these groups been assigned to view different treatment conditions. Further, the instability of the differences eliminates the possibility of developing "individual equations" to adjust the ratings of each group of observers to comparability. [pp. 29-30]

A third difficulty in obtaining reliable ratings of behavior over the course of an experiment, even within single subject designs, is observer drift (Johnson & Bolstad, 1970; O'Leary & Kent, 1973). Observers may possibly modify their application of the behavioral code over the course of a study. Some behaviors may be rated more strictly, others more leniently. Thus, the same behaviors that are recorded as present during the initial phases of the study may be recorded as absent toward the end of the study and vice versa.

O'Leary and Kent (1973) aptly stated the conclusions to be drawn from the current methodology research:

> We feel strongly that experimenters using group designs in field-experimental settings where small but significant differences have been found may have produced differences which are the result of methodological problems alone. In particular, we feel that the observer who has long been used as if he were a cumulative recorder must be viewed as a source of systematic variability which may greatly confound certain data. [p. 16]

Means of avoiding the problems of observational data are available and can be incorporated into the design of an experiment. However, procedures for avoiding all of the possible errors become complex, expensive, and impractical in many situations. For example, Johnson and

Bolstad (1973) suggested that certain methods of training observers, as well as the feedback given to them on their rating protocols as compared with a standard, may reduce or eliminate observer drift. Certain training procedures may possibly reduce errors of observation. Training observers in order to prevent errors appears to be much more practical than designing controls into an experiment.

Systematic training of observers may reduce possible observer errors. The use of consistent rating standards may lead to more consistent ratings by observers (Johnson & Bolstad, 1973; Kent, personal communication).

The present study involved the comparison of two types of training procedures: (1) training of observers by one individual and (2) self-training, or each pair of observers training itself. The consistency of the standard to which the raters were trained to conform decreased from the first to the second of these groups. After training was completed, observers recorded disruptive behaviors of children from videotapes of a nursery-school classroom.

Method

Subjects

Sixteen female undergraduate students were recruited from courses in Child Development at the University of North Carolina at Greensboro. Students were informed that they would serve as research assistants for a study involving nursery school children. Participation in this study fulfilled their course requirement for 10 hours of child observation. Selection of students for participation was primarily dependent on available time and lack of experience with formal observation of classroom behavior.

Trainer

The trainer was a female graduate student in psychology who had had 40 hours of experience with the disruptive behavior code.

Procedure

All subjects received a copy of the disruptive behavior code (O'Leary, Kaufman, Kass, & Drabman, 1970) prior to the beginning of training (see Appendix A). The nine categories of disruptive behavior were:

1. Out-of-chair: movement of the child from his chair when not permitted or requested by teacher. No part of the child's body is to be touching the chair.

2. Modified out-of-chair: movement of the child from his chair with some part of the body still touching the chair (excluding sitting on feet).
3. Touching others' property: child comes into contact with another's property without permission to do so. Includes grabbing, rearranging, destroying the property of another, and touching the desk of another.
4. Vocalization: any unpermitted audible behavior emanating from the mouth.
5. Playing: child uses his hands to play with his own or community property so that such behavior is incompatible with learning.
6. Orienting: the turning or orienting response is not rated unless the child is seated and the turn must be more than 90 degrees, using the desk as a reference point.
7. Noise: child creating any audible noise other than vocalization without permission.
8. Aggression: child makes movement toward another person to come in contact with him (exclude brushing against another).
9. Time off task: child does not do assigned work for entire 20-second interval. For example, child does not write or read when so assigned.

Subjects were also instructed on the method of reliability calculation.

Seven 60-minute training sessions were held to allow the subjects to familiarize themselves with and to practice the disruptive behavior code. Training sessions consisted of discussion of the categories as well as practice in using the code. All subjects viewed the same 40-minute videotapes of nursery-school children. A different tape was used each session.

Observations were made on a 20-second observe, 10-second record basis. A cassette recording synchronized with the videotapes marked the intervals. Any or all of

the nine categories of disruptive behavior could be recorded in any given interval. The same behavior could not be recorded more than once in any 20-second interval. Behaviors occurring during the 10-second interval were not recorded. If no instances of disruptive behavior, as defined by the above categories, occurred in a 20-second observation interval, observers recorded the category of "absence."

Observers were assigned to training groups on the basis of their available time. Pairs were assigned randomly within a condition.

Group I. The four pairs of observers in Group I were trained by the graduate student trainer. All questions about the code were answered by the trainer. In addition, observers compared their ratings with those of the trainer. The trainer's ratings and clarification of the categories were accepted as the standard.

Group II. The four pairs of observers in Group II were trained by themselves. The observers established their own interpretations of the code and set their own standards. The trainer was present during these sessions, but did not answer any questions concerning the code.

Data collection began after training was completed. Six 40-minute videotapes were rated by both groups. Each observation session was divided into four 10-minute segments. Two 10-minute blocks were used for within-pair reliability measures, and two blocks were used for

between-pair reliability assessments. Observers were informed of one of the within-pair and one of the between-pair reliability assessments. The order of informed and non-informed assessments, as well as between-pair and within-pair assessments were randomized. Thus, overtly and covertly assessed reliabilities were obtained for both within-pair and between-pair combinations.

Reliability

Reliability of observations was calculated within pairs, i.e., between members of an observation pair, and between pairs within each experimental group. The method of reliability calculation used was the number of agreements divided by the number of agreements plus disagreements. An agreement was scored if both observers recorded the same behavior in the same 20-second interval. A disagreement was scored if one observer recorded a behavior and the other did not. The category of "absence" was not included in the reliability calculations.

Results

The data for the six days of data collection were combined in two-day blocks. The data were combined in this manner in order to avoid the problems associated with the estimation of missing data. In addition, significant differences were not expected to occur between individual days. Rather, differences were expected to occur after several days.

A repeated measures analysis of variance with two between-group factors, type of training group (training by one trainer and self-training) and reliability assessor (within-pair and between-pair reliability assessment), and two within-subject factors, type of reliability assessment (overt and covert assessment) and time was performed on the reliabilities of the observers. Since the measure of reliability (percentage of agreement) is a proportion, an arcsin transformation was performed (Winer, 1971, pp. 399-400). Table 1 contains the means and standard deviations for the main factors.

The results of the repeated measures analysis are presented in Table 2. The analysis of variance did not reveal any differences in the means of the two training groups. Overtly assessed reliabilities were found to be significantly higher than covertly assessed reliabilities ($p < .05$). In addition, the within-observer pair reliabilities were found to be significantly higher than the

TABLE 1

Means and Standard Deviations of Reliabilities and Arcsin Transformed Reliabilities for Main Factors[1]

	Mean	Arcsin transformed mean	Standard deviation	Arcsin transformed standard deviation
One-trainer group	.6610	1.9135	.1190	.2711
Self-training group	.6606	1.9064	.1190	.2530
Overtly assessed reliabilities	.6838	1.9627	.1212	.2744
Covertly assessed reliabilities	.6379	1.8571	.1120	.2378
Within-pair reliabilities	.7044	2.0076	.1259	.2829
Between-pair reliabilities	.6173	1.8123	.0926	.1950
Days 1 and 2	.6050	1.7913	.1389	.2944
Days 3 and 4	.6844	1.9616	.0996	.2361
Days 5 and 6	.6931	1.9769	.0940	.2102

[1]See Appendix B for reliabilities and arcsin transformed reliabilities for individual pairs.

TABLE 2

Repeated Measures Analysis of Variance
on Arcsin Transformed Reliabilities

Source of Variance	df	SS	F
Between			
Type of Training Group (TT)	1	.0012	.0082
Reliability Assessor (RA)	1	.9156	6.1809*
TT X RA	1	.0376	.2536
S(TT X RA) error	12	1.7776	
Within			
Type of Reliability Assessment (TA)	1	.2676	6.6993*
TT X TA	1	.0077	.1938
TA X RA	1	.1764	4.4173
TT X TA X RA	1	.0028	.0713
TA X S(TT X RA) error	12	.4792	
Time (T)	2	.6790	16.2655**
TT X T	2	.0989	2.3692
RA X T	2	.0909	2.1782
TT X RA	2	.0347	.8318
S X T(TT X RA) error	24	.5010	
TA X T	2	.0400	.4237
TT X TA X T	2	.1004	1.0656
TA X RA X T	2	.0282	.2987
TT X TA X RA X T	2	.0934	.9911
TA X S X T(TT X RA) error	24	1.1314	

* $p < .05$

** $p < .01$

reliabilities between observer pairs ($p < .05$).

The change in reliabilities over time was found to be significant ($p < .01$). A Newman-Keuls post-hoc analysis indicated that the reliabilities for Days 3 and 4 and for Days 5 and 6 were significantly higher than the reliabilities for Days 1 and 2 ($p < .01$).

Differences that approach significance were found for the interaction between the type of reliability assessment and who the reliability assessor was ($p < .10$), with the overtly assessed within-observer pair reliabilities being higher than the covertly assessed within-pair reliabilities. This difference was greater than the difference between the overtly assessed between-observer pair reliabilities and the covertly assessed between-observer pair reliabilities.

In order to test the hypothesis that the groups were applying the disruptive behavior code differently, a repeated measures analysis of variance, as described above, was performed on the mean number of disruptive behaviors rated per interval. The group that was trained by one trainer recorded significantly more behaviors than did the self-training group ($F(1,28) = 34.4525$, $p < .01$).

In addition, significant differences were found in the variances of the mean number of disruptive behaviors rated per interval by the two training groups. The variability of the one-trainer group ($s^2 = .039$) was less than the variability of the self-training group ($s^2 = .068$) ($F(95,95) = 1.74$, $p < .01$).

Discussion

The results of the present study corroborate the existence of errors associated with the use of human observers in the collection of observational data. The finding that overtly assessed reliabilities are higher than covertly assessed reliabilities lends additional support to the findings of Reid (1970) and Romanczyk et al. (1973). The present results taken in conjunction with the findings of Johnson and Bolstad (1973) and O'Leary and Kent (1973) add to the accumulating evidence against the use of occasional overt reliability assessments in behavioral research. It appears that the reliabilities reported from samples of overt assessments are considerably higher than the reliabilities found under conditions of covert assessment. Occasional overt reliability assessments in behavioral research appear to be poor measures of the overall reliabilities of observers. Since the conditions for overt assessment present the observer with a different stimulus situation than the conditions for covert assessment, the differences obtained are not surprising.

The finding that within-observer pair reliabilities were higher than between-observer pair reliabilities, even when the same training procedures were used, adds an additional challenge to the validity of observational data. The

training procedure allowed for group discussion of the categories, thus increasing the probability of consensus among all members of a training group. Even with this advantage, observers apparently adopt idiosyncratic interpretations of the code and conform more to the rating standards of their partner than to the standards of others in the group. This result has strong implications for studies which utilize multiple observers, such as different observers in different classrooms or different observers in the same classroom on different days of the week.

The change of reliabilities over time was significant even in the brief data collection period of ten days. Although it was predicted that the reliabilities would decrease with time, the results indicated that reliability increased. This increase may have been an artifact of the time constraints on the study. The 7-session training phase was not long enough for the reliabilities to reach asymptote and to stabilize. Thus, observers were continuing to become more proficient with the code during the data collection phase. If the training phase had been continued until the reliabilities became stable, it is possible that a decrease in reliabilities would have been found during the data collection phase.

The finding that reliabilities were not stable over time implies that studies using observers are obtaining data with varying reliability. In addition, the change in reliabilities is likely to be accompanied by changes in the

interpretation of the code. Thus, behaviors recorded in the early phases of a study may not be comparable to behaviors recorded in the later phases of the study. The change in reliabilities over time may be interpreted as evidence of observer drift. These results, taken in conjunction with the findings of Johnson and Bolstad (1973) and O'Leary and Kent (1973) suggest that investigators must be cautious of drawing conclusions based on small but significant results in studies utilizing observational data.

Although the analysis of variance on the reliabilities did not reveal significant differences between the two training procedures, other differences were found which lend support to Johnson and Bolstad's (1973) suggestion that the use of certain training techniques may increase the validity of observational data. The analysis of variance on the mean number of disruptive behaviors rated per interval indicated that Group I rated significantly more behaviors than Group II. This result suggests that the type of standard used for training may determine how the code is interpreted, and therefore which behaviors are recorded. Thus, in studies where observers are trained informally or in which different observers or groups of observers are trained by different standards, data may be collected in which different observers record different behaviors.

In addition, it was found that the variability of the mean number of disruptive behaviors recorded per interval was affected by the type of training procedure. The variance of the mean number of disruptive behaviors recorded by Group I, which was trained by a consistent standard, was significantly smaller than the variance of Group II. Thus, an inverse relationship was found between the number of disruptive behaviors recorded and the variance of the mean number of disruptive behaviors recorded per interval.

The finding that the variance of the mean number of disruptive behaviors recorded per interval for Group I was smaller than the variance of Group II suggests that observers trained by a consistent standard produce more stable recording of behavior. In addition, although no differences were found in the reliabilities of the two training groups, it would appear that it would be more difficult for Group I to achieve reliabilities similar to those of Group II since Group I rated significantly more behaviors than did Group II. The finding of an inverse relationship between the number of behaviors recorded per interval and the variance of the mean number of behaviors recorded per interval, even with similar reliabilities, suggests that measures of reliability alone may not be adequate for investigators to assume that observers are rating consistently.

In addition, the greater stability in the mean number of behaviors rated per interval in conjunction with the greater number of behaviors recorded per interval by the one-trainer group raises the question of accuracy of recording. This finding lends additional support to the above statements that stress the importance of training all observers in a systematic manner. Casual interpretations of the code should be avoided. Observers should be instructed to consult one trainer, or standard, rather than consulting each other when questions concerning the observational code arise.

The results of the present study have only begun to present solutions to the problems associated with observational data. The present research could be refined and expanded in several ways. Future research should include observers being trained to a given criterion of agreement before data collection is begun. In addition, observers should be trained until their reliabilities reach an asymptote. These modifications would permit the investigator to draw more conclusive interpretations of the effect of time on observer reliability.

References

Arrington, R. E. Interrelations in the behavior of young children. Child Development Monographs, Monograph No. 8. New York: Teachers College, Columbia University, 1932.

Johnson, S. M. & Bolstad, O. D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Ill.: Research Press, 1973.

Jones, R. R. Behavioral observation and frequency data: Problems in scoring, analysis and interpretation. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Ill.: Research Press, 1973.

Kass, R. E. & O'Leary, K. D. The effects of observer bias in field-experimental settings. Paper presented at a symposium entitled "Behavior analysis in education," University of Kansas, Lawrence, April 1970.

Loomis, A. M. A technique for observing the social behavior of nursery school children. Child Development Monographs, Monograph No. 5. New York: Teachers College, Columbia University, 1931.

O'Leary, K. D., Kaufman, K. F., Kass, R. E., & Drabman, R. S. The effects of loud and soft reprimands on the behavior of disruptive students. Exceptional Children, 1970, 37, 145-155.

O'Leary, K. D. & Kent, R. Behavior modification for social action: Research tactics and problems. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Ill.: Research Press, 1973.

Olson, W. C. The measurement of nervous habits in normal children. University of Minnesota, The Institute of Child Welfare, Monograph Series No. III. Minneapolis: The University of Minnesota Press, 1929.

Reid, J. B. Reliability assessment of observation data: A possible methodological problem. Child Development, 1970, 41, 1143-1150.

Romanczyk, R. G., Kent, R. N., Diament, C., & O'Leary, K. D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Skinrud, K. Field evaluation of observer bias under overt and covert monitoring. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior change: Methodology concepts and practice. Champaign, Ill.: Research Press, 1973.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Appendix A

Disruptive Behavior Code--Child

1) Out of Chair--symbol = O

Purpose: Out of chair is intended to monitor the gross motor behavior of the child removing himself from his seat entirely. Such behavior (when not permitted) may interfere with the child's learning and is potentially distracting to others, e.g., running around the room.

Description: Observable movement of the child from his chair when not permitted or requested by teacher. None of the child's weight is to be supported by the chair, but the child may be in physical contact with the chair.

Critical Points: None of the child's weight is to be supported by the chair.

Includes: Child is leaning on desk and has either lost all contact with the chair or none of his weight is actually being supported by the chair. Time limits on the following beginning with the teacher's permission: Allow 15 seconds for a child to get from the teacher's desk to his own. Allow 15 seconds for a child to return to his own seat after completing a task (e.g., placing a word card on the wall). Pencil sharpening--1 1/2 mins. Getting a drink--1 1/2 mins. (fountain in room). Getting a book--1 1/2 mins. (time limit starts from the second that the child gets

out of seat). Going to the bathroom: (a) 2 min. limit, (b) 30 sec. limit beginning when child leaves bathroom.

Note: If the child returns to the chair after 1 1/2 (or 2 mins., where applicable), but during the 10 sec. inter-interval period, the "0" will be recorded in the 20 sec. interval just prior to the 10 sec. interval.

Going to get a reading book during a math lesson. When child is fully standing and the back of the legs touch chair, or child is fully standing and is touching back of chair with hands. Going to the teacher's desk when not permitted. Throwing away papers. Stretching (if child actually leaves seat).

Excludes: Retrieval of an accidentally dropped task-related object. Leaning forward to pick up an object even if all contact with the chair is momentarily lost, providing the child is not standing fully erect on feet. Include if child begins crawling around on floor after retrieving object. Also include if child is moving from desk in a crouched position, so as not to let the teacher see him, etc.

2) Modified Out of Chair--symbol = Ø

Purpose: Modified out of chair is intended to monitor less intense motor behavior than displayed in out of chair, and behavior which is usually only distracting

for the child himself rather than others.

Description: Movement of child from his chair, with some of his weight still being supported by the chair.

Critical Points: The child is still at his desk and some of his weight is being supported by the chair.

Includes: Leaning forward to pick up an object even if all contact with the chair is momentarily lost, providing the child is not standing fully erect on feet. Bouncing in chair, e.g., in responding excitedly to some event. Kneeling on chair. Sitting on back of chair. Both feet on or in desk. Lying across chair horizontally. Standing near desk with one foot on the chair.

Excludes: When child is fully standing and the back of legs touch chair. Sitting on one or both feet. One "cheek" off chair.

3) Touching Others' Property--symbol = T

Purpose: Touching is intended to monitor behavior which is distracting to the child and very often to others when the child comes into contact with the personal property of another.

Description: Child comes into contact with another's property without permission to do so.

Critical Points: The child does not have permission for his action and not that his action may or may not result in an alteration and post hoc permission.

Includes: Grabbing, re-arranging, destroying the property of another. Using material object as extension of hand to touch others' property. Hand brushing on others' desk if this act is incompatible with learning (i.e., the child is attending to the act). Touching desk of another, whether other person is seated in it or not (this includes teacher's desk). Resting elbows on desk behind if this act is incompatible with learning or annoys the other child.

Excludes: Touching others on the back or any part of the body or clothing. Use of shared possessions such as rulers, erasers, art materials. Elbow resting on another's desk or hand brushing against it, if the desks are together and neighbor is not disturbed and such an act is not incompatible with learning. Walking past a desk, chair, etc., and accidentally brushing or touching the desk, chair, etc., i.e., child is not attending to the behavior.

Note: When child is at teacher's desk with permission, and is waiting to be helped, do not score idle touching of objects on teacher's desk. Touching should be scored, if the teacher specifically instructs child to stop and child continues or if child is instructed to perform some task at desk and then begins to touch objects on desk.

4) Vocalization--symbol = V

Purpose: Vocalization is intended to monitor verbal behavior which is usually distracting to both the child and to others.

Description: For the sake of consistency, any audible non-permitted vocalization is to be recorded even though in the opinion of the observer it did not "seem" disruptive. Any non-permitted audible behavior emanating from the mouth.

Critical Points: The observer must actually hear the vocalization. Inferences are not acceptable except as noted below.

Includes: If vocalization is obvious, but can't be heard (obvious--if another child responds). Answering without being called on. Moaning. Yawning. Any noise made with the mouth when eating--unless the child has permission to eat. Any vocalization made in response to the disruptive behavior of another child, e.g., telling another child to return stolen article, crying in response to aggression committed to his person or possessions, etc., if the child has not received permission specifically from the teacher to speak. Whispering. Belching. Crying. Shouting. "Operant" coughs or sneezes.

Excludes: Vocalization in response to teacher's question. Sneezing. Automatic coughing.

Note: Once a child is recognized by the teacher vocalization is not scored, regardless of the content of the vocalization: crying, yelling, swearing, etc., until the teacher specifically instructs the child to stop.

5) Playing--symbol = P

Purpose: Playing is intended to monitor often subtle manipulative behavior that is distracting to the child and possibly also distracting to others.

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

Critical Points: Child uses his hands to manipulate his own or community property.

Includes: Playing with toy car when assignment is spelling. Playing with comb or pocket book. Eating only when the hands are being used--chewing gum is not rated as P unless child touches or manipulates it with his hands. Poking holes in workbook. Cleaning nails with pencil. Drawing on self. Manipulating pencil in such a manner as to make the behavior incompatible with learning, e.g., shoving pencil back and forth on desk; waving pencil through air as an airplane. Picking scabs, nails, or nose if the desired "object" is separated from the body and manipulated. Looking into desk and moving arms, but does not come out with task-related

object. Working with or reading non-task-related material, e.g., reading page 25 when told to read page 1, doing math when told to do spelling, etc.

Excludes: Touching others' property. Playing with own clothes.

Note: Include if article is removed from body, e.g., shoes, tie, buttons, scarf, etc., and is manipulated.

Lifting desk or chair with feet (rate N if this creates audible noise). Random banging of pencil on desk (rate N, if audible). Simple twiddling of pencil, if it is not seen as being incompatible with learning.

Note: Rate twiddling of pencil, banging pencil, or putting pencil in mouth, hair, behind ear, etc., if child attends to such behavior and ceases attending to assigned task. Operational definition of attending: child either looks at manipulated object or begins to manipulate object in non-random patterns for more than 5 seconds.

Picking scabs, nails, or nose if the desired "object" is not separate from the body.

6) Orienting Response--symbol =)

Purpose: Orienting is intended to monitor the gross motor behavior of turning around from the designated point of reference. Such behavior is distracting to child since it usually precludes attending to assigned

task, and is often distracting to others.

Description: Child turns more than 90 degrees from point of reference while seated.

Critical Points: The child must be in his seat; he may be in a modified position; and orienting includes both the horizontal and vertical axis.

Includes: Turning to the person behind. Looking to the rear of the room. Turning around in chair or turning chair around. Leaning back in chair more than 90 degrees.

Note: Point of reference is typically child's desk, but may be the teacher if the children are directed to attend to her. If child should turn desk at some angle, point of reference becomes where desk was originally, not to where the child has moved it. Also, the child's chin should be used as the indicator of how far he has turned. Therefore, orienting is rated when the child's chin has turned more than 90 degrees from point of reference.

Excludes: Orienting during class discussion when the teacher directs (either implicitly or explicitly) the class to attend to a child's explication of an answer. Orienting while picking up a task related object. When child is in corner or otherwise out of his chair.

7) Noise--symbol = N

Purpose: Noise is intended to monitor the frequency of

of distracting sounds produced by the child, other than vocalization.

Description: Child is creating any audible noise, without permission, other than vocalization. For the sake of consistency, any audible sound is to be recorded even though in the observer's opinion, it did not "seem" disruptive.

Critical Points: The observer must actually hear the sound to rate it. Inferences are not acceptable.

Includes: Turning pages in an exaggerated manner, producing noise. Moving desk around. Pencil tapping. Banging of any object. Fishing in desk without coming out with anything or coming out with an inappropriate object (if noise is actually made in the process). Shuffling feet more than once each way. Any noise made while getting out of chair without permission. In general, any noise made in conjunction with any disruptive behavior, e.g., any noise made when the child throws a book or other object at another.

Excludes: Shuffling feet (if only once each way). Accidental dropping of a task-related object (book or pencil). Pushing chair back and forth once during a permitted act (e.g., to get a task-related object).

8) Aggression--symbol = Ag

Purpose: To measure the highly disruptive behavior of physical assaults.

Description: Child makes an intense movement directed at another person so as to come into contact with him, either directly or by using a material object as an extension of the hand.

Critical Points: Intention is to be recorded rather than just accuracy of assault. (E.g., aggression is recorded if child throws pencil or swings at another, regardless of whether or not the pencil or motion hits the child.)

Includes: Blocking others with arms or body from attaining goal (e.g., while walking up aisle). Tripping. Kicking. Throwing.

Excludes: Brushing against another (include if action is continually repeated so as to tease or annoy).

9) Time-Off-Task--symbol = X

Purpose: Time-off-task is intended to monitor non-attending behavior, that, if excessive, is detrimental to child's performance.

Description: Child does not do assigned work for entire 20 second interval.

Critical Points: Child makes no attending response for the entire 20 second interval. Child must only attend, i.e., "look at," his work. Inferences that "he isn't really thinking about it" are not acceptable.

Includes: Child does not write when assigned to do so. Child does not read when so assigned. Child is working on inappropriate material, e.g., on math during spelling,

etc. Daydreaming--as reflected in not working. Child does not ask teacher for additional work or help when finished with assigned task, and merely sits at desk or begins to play for entire interval. When in corner, child's head must be within a 45 degree angle from the corner formed by two walls (i.e., if his head is facing either of the two walls directly, for a 20 second period, he would be rated X).

Excludes: Child has his hand raised to ask questions. Child is told he may cease working if he so desires.

10) No inappropriate behavior as defined by the above categories--symbol = Ab

Appendix B

Reliabilities and Arcsin Transformed Reliabilities for Individual Pairs

		Days 1 and 2 Assessment		Days 3 and 4 Assessment		Days 5 and 6 Assessment	
		Overt	Covert	Overt	Covert	Overt	Covert
Group I							
Within-Pair Reliabilities							
Subjects 1 & 2	R^2	.80	.85	.95	.78	.87	.76
	T^3	2.2143	2.3462	2.6906	2.1652	2.4039	2.1176
Subjects 3 & 4	R	.82	.69	.91	.77	.83	.79
	T	2.2653	1.9606	2.5322	2.1412	2.2916	2.1895
Subjects 5 & 6	R	.74	.61	.53	.63	.61	.57
	T	2.0714	1.7926	1.6308	1.8338	1.7926	1.7113
Subjects 7 & 8	R	.65	.41	.67	.70	.65	.49
	T	1.8755	1.3898	1.9177	1.9823	1.8755	1.5508
Between-Pair Reliabilities							
Subjects 1 & 3	R	.63	.57	.68	.65	.59	.77
	T	1.8338	1.7113	1.9391	1.8755	1.7518	2.1412
Subjects 2 & 6	R	.60	.54	.59	.57	.71	.67
	T	1.7722	1.6509	1.7518	1.7113	2.0042	1.9177
Subjects 4 & 7	R	.48	.52	.58	.62	.63	.62
	T	1.5308	1.6108	1.7315	1.8132	1.8338	1.8132
Subjects 5 & 8	R	.62	.46	.73	.57	.63	.62
	T	1.8132	1.4907	2.0488	1.7113	1.8338	1.8132

Reliabilities and Arcsin Transformed Reliabilities for Individual Pairs (continued)

		Days 1 and 2 Assessment		Days 3 and 4 Assessment		Days 5 and 6 Assessment	
		Overt	Covert	Overt	Covert	Overt	Covert
Group II							
Within-Pair Reliabilities							
Subjects 9 & 10	R	.71	.61	.74	.73	.81	.61
	T	2.0042	1.7926	2.0714	2.0488	2.2395	1.7926
Subjects 11 & 12	R	.80	.76	.77	.75	.71	.80
	T	2.2143	2.1176	2.1412	2.0944	2.0042	2.2143
Subjects 13 & 14	R	.78	.27	.67	.64	.78	.73
	T	2.1652	1.0928	1.9177	1.8546	2.1652	2.0488
Subjects 15 & 16	R	.53	.63	.74	.69	.83	.64
	T	1.6308	1.8338	2.0714	1.9606	2.2916	1.8546
Between-Pair Reliabilities							
Subjects 9 & 12	R	.44	.65	.71	.68	.63	.66
	T	1.4505	1.8755	2.0042	1.9391	1.8338	1.8965
Subjects 10 & 14	R	.44	.60	.51	.81	.84	.61
	T	1.4505	1.7722	1.5908	2.2395	2.3186	1.7926
Subjects 11 & 15	R	.44	.57	.67	.68	.73	.66
	T	1.4505	1.7113	1.9177	1.9391	2.0488	1.8965
Subjects 13 & 16	R	.69	.45	.62	.56	.73	.60
	T	1.9606	1.4706	1.8132	1.6911	2.0488	1.7722

[2]R = Reliability Coefficients

[3]T = Arcsin Transformed Reliability Coefficients