8-1974

The Effects of Observer Load on Inter-Observer Agreement

Victor A. Dotson

Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses

Part of the Psychoanalysis and Psychotherapy Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2502

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
THE EFFECTS OF OBSERVER LOAD
ON INTER-OBSERVER AGREEMENT

by
Victor A. Dotson

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment
of the
Degree of Master of Arts

Western Michigan University
Kalamazoo, Michigan
August 1974
ACKNOWLEDGEMENTS

I wish to thank all those who have assisted and encouraged me in the preparation of this. Professors John Michael and Brad Huitema provided valuable insight and fair criticism. A great debt is owed Professor Robert P. Hawkins for his persistence and vigilance in producing the clearest possible manuscript. I thank Karen Williams for an excellent typing job and my wife, Claudia who can't be thanked enough. To all these and to the many others at Western Michigan University, who have given much needed help, I can only express my gratitude and appreciation for their assistance.

Victor A. Dotson
INFORMATION TO USERS

This material was produced from a microfilm copy of the original document. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the original submitted.

The following explanation of techniques is provided to help you understand markings or patterns which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting thru an image and duplicating adjacent pages to insure you complete continuity.

2. When an image on the film is obliterated with a large round black mark, it is an indication that the photographer suspected that the copy may have moved during exposure and thus cause a blurred image. You will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., was part of the material being photographed the photographer followed a definite method in “sectioning” the material. It is customary to begin photoing at the upper left hand corner of a large sheet and to continue photoing from left to right in equal sections with a small overlap. If necessary, sectioning is continued again — beginning below the first row and continuing on until complete.

4. The majority of users indicate that the textual content is of greatest value, however, a somewhat higher quality reproduction could be made from “photographs” if essential to the understanding of the dissertation. Silver prints of “photographs” may be ordered at additional charge by writing the Order Department, giving the catalog number, title, author and specific pages you wish reproduced.

5. PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Xerox University Microfilms
300 North Zeeb Road
Ann Arbor, Michigan 48106

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
MASTERS THESIS M-6323

DOTSON, Victor A.
THE EFFECTS OF OBSERVER LOAD ON INTER-OBSERVER AGREEMENT.

Western Michigan University, M.A., 1974
Psychology, clinical

Xerox University Microfilms, Ann Arbor, Michigan 48106

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>METHOD</td>
<td>9</td>
</tr>
<tr>
<td>RESULTS AND DISCUSSION</td>
<td>19</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>64</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>66</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>69</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>70</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
It is generally agreed that a major responsibility of a science is to assure the accuracy of its observations. This task is especially difficult in the science of Applied Behavior Analysis. Because of the emphasis on naturalistic settings, human observers are the principal means of data collection. In employing human recorders, great effort must be taken so that the observers are highly accurate and do not incorporate their own or the experimenter's bias into the data. Logically, inaccuracy between observers may result from several causes: (1) the observers may be inadequately trained, (2) the task may be overly difficult, (3) the definitions of behaviors from which the observers draw their observational criteria may be inadequate and, (4) as responding organisms the observers are affected by factors operating in the immediate environment, such as the likelihood of the data being checked for accuracy or the presence of the experimenter or other supervisory personnel.

Some indication of the relevant variables affecting observer agreement have been noted in the Applied Behavior Analysis literature. Bijou, Peterson, and Ault (1968) wrote that disagreements between observers may be related to inadequacies in the observational code, the training of observers, or the method of calculating the agreement scores. Romanczyk, Kent, Kiament, and O'Leary (1971)
demonstrated that agreement scores would fluctuate depending on which "reliability assessor" observers were paired with. An artificial difference in the scoring of four behavior categories was established between two reliability assessors and a twenty per cent difference in agreement resulted when observers were informed of which assessor was computing the reliability. Reid (1970) concluded that reliability suffers greatly when it is either never checked (subsequent to the beginning of a study) or checked infrequently and only when the observers are notified (spot-checking). O'Leary and Kent (1973) observed changes in observer agreement in relation to several different variables. It was found that reliability scores suffered when covert assessments of agreement were made, when a supervising experimenter was present, and when the scores were calculated by the experimenter rather than the observers. Hawkins and Dotson (1972), in discussing interval recording (see Bijou et al., 1968), made a clear distinction between the three types of agreement calculations commonly observed in behavior analysis literature and graphically illustrated that the most commonly used index may significantly over-estimate observer agreement.

One variable not yet investigated is the effect of "observer load" upon agreement scores. Observer load is the number of behavior classes which the observer must
concurrently remember, discriminate among, and record. Many studies have been reported which use an extensive observational code to classify a wide variety of subject and environmental events. For example, Madsen, Becker, and Thomas (1968) employed a code categorizing 31 events and behaviors. Ray, Shaw, and Patterson (1968) formulated a school observation technique incorporating twenty-nine different symbols which could be combined in innumerable ways. This system was later revised by Cobb and Ray (1971) into a twenty-six category code. Walker and Buckley (1972) used a twenty category system in their research. Given the previously noted manipulability of observational agreement scores, the reliability and generality of those studies which use a large number of categories must be questioned.

For example, a case might easily be made that experiments which employ a large number of categories are overweighted with low rate behaviors. The necessarily high average reliability produced by this (see Hawkins and Dotson, 1972) would overshadow important disagreements between observers on other categories. Hence, the conclusions professed by the experimenters and predicated on highly reliable data may not hold true for all behaviors examined.

The Madsen, et al. (1968) study on rules, praise and ignoring in the elementary classroom provides a good
illustration of what may be heavily loading an observer. The observers scored either ten categories of child behaviors or twenty-one definitions of behavior classes, observed the subject, monitored a stopwatch, and recorded the appropriate symbol in the appropriate space on the data sheet. Although the authors report high average reliability over subjects, behavior classes, and days, their basic reliability procedure was by spot-check. O'Leary and Kent (1973) have shown, if the observers are aware of their accuracies being checked, the agreement score may peak on those occasions, only to decrease on unchecked days. If inter-observer agreement was taken only on one behavior class at a time and the observers were also aware of this, they may have concentrated more on the checked behavior category and the recording of other categories may have lagged. Also, it is not clear exactly what technique Madsen, et al. used to calculate the reliability. Their description is insufficient for differentiating between the "interval-by-interval" method and the "scored-interval" method as described by Hawkins and Dotson (1972). High reliabilities may be produced solely by the rate of behavior, if the "interval-by-interval" reliability score is used.

A secondary purpose for the origination of this research was to determine the relationship between some of the more commonly used calculations of observer
agreement. Bijou et al. (1968) alluded to each of the four calculations employed in this study. The frequency calculation (F) consists of the ratio of one observer's frequency data to another. The smaller total is divided by the larger to produce a ratio of 1.0 or below. As noted by Bijou et al. this procedure does not indicate whether the two observers were simultaneously recording an event i.e., one of the observers could record many behaviors during the first half of a session and few during the second half, if the second observer scored many behaviors during the second half of the session but indicated a low rate of the behavior during the session's early part, then it would occur that a high ratio of frequencies would be produced with substantially little agreement between observers.

By segmenting the session into small equally sized sections (usually 10 second blocks) and determining the agreement between observers for each segment, the problems of accidental similarity of totals may be averted. This has been the rationale for using interval recording systems and the interval-by-interval agreement calculation (I-I). Procedurally this calculation involves comparing simultaneous intervals of the data sheets of two observers and scoring each interval as an agreement or disagreement between those observers. An agreement is defined as both observers scoring the same symbol or sets
of symbols (representing a defined response) within an interval; whereas a disagreement is defined as one of the observers marking a different symbol in an interval when compared to another observer. The total number of agreements are then divided by the total number of intervals in the session.

The I-I calculation, however, is not without its limitations. As Bijou et al. noted and Hawkins and Dotson (1972) analyzed in detail, when rates of behavior are low there will of necessity be a great number of intervals where the observers agree only on the nonoccurrence of the behavior. The resulting ratio could be very high even though no single agreement on the occurrence of the behavior was scored. Bijou et al. pointed out that two calculations of agreement could be computed, one for the occurrence and one for the nonoccurrence of the behavior. No procedures for the calculations were presented, however. Hawkins and Dotson formalized these calculations and renamed them. The calculation of agreement on occurrences became Scored Interval (SI) agreement and Unscored Interval (UI) was substituted for the nonoccurrence agreement score.

To obtain the SI agreement score, only those intervals where one or both of the observers record the behavior as occurring are considered. These intervals are compared and scored as agreements or disagreements. The number of
agreements are then divided by the number of agreements plus the number of disagreements. Those intervals in which the observers agree on the nonoccurrence of the behavior are deleted for the purposes of this calculation. Conversely the UI agreement score is obtained by deleting those intervals where the observers agree on the occurrence of the behavior. The remaining intervals are scored as match or mismatch and the agreements are then divided by the number of agreements plus disagreements.

A method for the interpretation of these two scores has not been devised. In cases where the two scores differ appreciably, no procedure exists to determine whether or not the data are acceptably reliable. In the opinion of the author the SI score should carry more weight because it focuses on the occurrence of the behavior.

The interrelationship between the three interval agreement calculations (I-I, SI and UI) is complex. When rates of observed behavior are low the UI score will resemble the I-I score without fail and the SI score will range independently of these two. When rates of observed behavior are higher (75-100 per cent of intervals scored), then the SI and I-I scores will be in close agreement with UI scores varying independently. As Hawkins and Dotson (1972) noted, only when the rate of observed behavior is precisely at 50 per cent of intervals rated may the I-I score drop all the way to zero per cent agreement.
Similarly, low frequencies of response necessitate high rates of UI agreement and high frequencies of response preclude low rates of SI agreement. By determining the average number of intervals recorded for a pair of observers, the precise range for each of the interval agreement scores is determined (see Hawkins and Dotson, 1972).

Any discussion of agreement scores would be insufficient without an adequate consideration of the purposes of reporting agreement scores. Hawkins and Dotson (1972) reported that the purpose of collecting agreement scores was essentially three-fold. First, it should tell the experimenter whether his definition is sufficiently clear, inclusive, and specific to produce objective data. Second, it should indicate whether the observers are performing competently. And, finally, it should determine the presence or absence of observer bias, i.e., it should indicate whether the experimental effect is believable. This last of course is the most important purpose since systematic errors in recording would seriously damage the utility of research results. Random observer errors produced by whatever means would not be quite as damaging but could greatly reduce the ability of the data to indicate changes in the behavior of the subject produced by manipulation of the independent variable. These considerations suggested a need for empirical evaluation
of the effects of different sized observer loads on each of the four cited methods of agreement calculation.

Method

Subjects

Four psychology graduate students served as the subjects (observers) for this experiment. They were selected from volunteers electing a course on behavior modification in a school setting. As part of this course they received instruction in a wide variety of methods for observing behavior in the classroom. In exchange for acting as observers for the experiment, they received partial course credit. None of the four subjects was informed of the purpose of the experiment, and they were urged not to discuss the proceedings among themselves. It appeared that they followed this instruction.

Apparatus

A Sony VTR system (Model AV3650) was used to record and play back videotapes obtained within a special education elementary classroom. The class consisted of ten children aged 9 through 12 years. A special education instructor and an aide worked with the children, who were considered to have serious school adjustment problems. The tapes were observed on a standard television set
(21-inch screen) modified to monitor videotapes. The television was placed directly on the floor and angled slightly upward toward the observers. A table which fit over the television held the video recorder. Sound from the recordings was provided through stereo earphones modified so that both channels carried the same signal.

Two of the subjects viewed the videotapes simultaneously. Each member of the pair was supplied with a switch box containing 17 SPST slide switches. The switch boxes were connected to a pair of interlocked Esterline-Angus twenty-pen event recorders. Both recorders were driven by the same drive mechanism to assure precisely the same drive speeds. The observers marked the onset and offset of any of seventeen behavior categories by actuating the corresponding switch. Behaviors were recorded on Esterline-Angus chart paper (Chart No. 1720 B). The event recorders were used rather than the standard arrangement (clipboard, stopwatch, paper, and pencil) because the event recorders would allow for continuous recording of all behaviors. This was necessary for the calculation of the Frequency agreement score (F). Demarcations in the chart paper made possible the imposition of an interval recording system at a later date. Occurrences within intervals were scored in the usual manner, i.e. only the first instance of the response within an interval
marked it as an "occurrence interval," if the response ended and began again within that interval it added nothing. It seemed likely that the continuous recording of behaviors might increase the load to observers since under the standard interval system the observers may discontinue vigilance for a response once it has been recorded within the interval. Consequently, if the response occurs earlier in the interval the observer's task in regard to that particular behavior is effectively finished until the beginning of the following interval.

In order to prevent the observers from on the action of the event recorders, a partition was mounted vertically on the table holding the video recorder. The partition also blocked the observers from viewing the experimenter who was seated behind the table and partition monitoring the operation of the recording equipment.

Independence between observers was assured by means of partitions, earphones, and background noise (see Figure 3). No observer could see what behaviors the other was recording due to the heavy cardboard partition separating them. Although the earphones alone may have been sufficient in eliminating sound cues produced by the event recorders, background clicking noises were also supplied through a series of electromechanical timers and counters to eliminate the possibility that one observer could tell when another recorded a response by the sound of the event recorder.
Variables

The independent variable was the number of behavior classes scored simultaneously. Six levels of this variable were employed: one, four, seven, ten, thirteen, and sixteen categories of behavior recorded at the same time.

The dependent variable was inter-observer agreement as calculated through a variety of measures. The data made by each pair of observers were segmented into ten-second intervals and manually scored for agreement in each interval (Bijou, Peterson, and Ault, 1969). Three calculations of interval observer agreement were obtained, interval-by-interval (I-I), scored interval (SI), and unscored interval (UI). A fourth calculation, frequency agreement (F), was produced by disregarding the interval demarcations and merely totaling the number of recorded onsets of a behavior for each observer, and dividing the smaller total by the larger. The definition of agreement was different for each of the four calculations, but each resulted in a ratio score which was translated to a percentage equivalent.

This number of dependent measures seemed necessary for several reasons. Except for the unscored interval (UI) calculation, all had been reported in use in a wide variety of contexts in the behavior analysis literature.
Although the calculations are vastly different, the scores which result are used interchangeably to represent the same variable, reliability of the data. Since it was felt that these scores were not equal in ability to represent reliability, and since it appeared likely that they might measure different aspects of observer agreement, all four calculations were used. In this way it was felt that the data from this experiment might provide a good means for comparing each of the calculations with the others and thus extend the utility of the data.

Observer Training

The observers were paired previous to the first training session and this arrangement maintained throughout the experiment. Each observer signed a contract pledging to be vigilant in his observations and to refrain from discussing the experiment (see Appendix A). At the first training session the observers then received a list of definitions for the seventeen behavior classes used in the experiment. The behavior definitions employed in this experiment came from a variety of sources. Many were borrowed verbatim from previous research on subjects in a classroom setting. The study by Madsen et al. (1968) and observational code developed by Ray, Shaw, and Patterson (1968) were the heaviest contributors. Some of the definitions were developed after
observing the behaviors of the subject. Many different types of definitions and behaviors were included on the principle of using a code which contained a great deal of diversity in likely rates of responses as well as adequacy of definitions. The titles of the responses recorded were Mouthing Objects (MO), Bad Posture (BP), Initiation to Peer (IP), Peer Initiation (PI), Work (W), Not Attending (NA), Sitting at Table (ST), Nodding Head (NH), Proximity (P), Recitation (R), Picking Objects off Floor (PF), Normative Behavior (NB), Inappropriate Fine Motor (IFM), Verbalization (VE), Turning Around (TA), Volunteers (VO), and Physical Contact (PC). The experimenter urged the observers to study the code diligently as they would be tested during the next training session.

During the next session the observers had a brief period to review the definitions. The experimenter then removed their lists and asked the observers to write down the definition for each behavior class after the experimenter recited its symbol. Definitions were checked by the experimenter and unacceptable accounts of the definitions required that the observer further study the definitions in question. In order to be acceptable, the definitions had to be nearly verbatim, but some substitutions of non-critical words or phrases were allowed.

Next, the observers were acquainted with the switch boxes and their operation. They then proceeded to practice
scoring of the seventeen behavior categories. Behaviors were recorded singly for ten minutes each. The experimenter first reviewed the definition of the behavior class, then started the videotape player and instructed the observers to begin recording. Following the practice scoring, the experimenter calculated the scored interval agreement and any recorded behavior not meeting a criterion of .70 was recorded again. If, after the second trial, criterion agreement was not obtained, the experimenter consulted with observers to resolve recording differences. The observers then recorded under close scrutiny from the experimenter the troublesome behaviors for five-minute periods. When the agreement scores approximated 70 per cent (SI), the observers attempted another ten-minute recording. This procedure was distributed over five training sessions for both pairs of observers and resulted in an agreement score of at least 70 per cent (SI) for ten-minutes for each of the seventeen behavior classes.

This method of observer scoring was employed in order to reduce the effects of implicit definitions produced as a result of observers comparing scores between themselves rather than with the experimenter (see Hawkins and Dobes, 1973).
Experimental Design

A single organism design was employed, in which each observer pair served as subjects under all six experimental conditions (number of behaviors recorded). In order not to confound effects of order of presentation, two schedules (A and B, see Appendices B and C) for presenting the six levels of independent variable were developed. Partial balancing of the six levels and seventeen categories was accomplished by sectioning the levels into three conditions, small (one and four behaviors), medium (seven and ten behaviors), and large (thirteen and sixteen behaviors). A second possible confounding could arise from the fact that certain behaviors would tend to produce lower agreement scores than others (due to their topography, social context, frequency of occurrence, etc.), so it was considered necessary to employ each definition approximately the same number of times in each experimental condition. In practice each definition appeared from one to three times in the small condition (one or four), and six to nine times in the medium condition (seven or ten), and nine to eleven times in the large condition (thirteen or sixteen).

Complete balancing of the design was deemed unfeasible since it would have required sixteen sessions in addition to training time. Under these taxing conditions
it appeared likely that the motivation for accurate performance would not have remained constant throughout, since the experiment would have extended beyond the end of the semester in which the observers were to receive credit. The resulting partially balanced schedules appeared to be adequate compromises.

Procedure

The pair of observers who followed Schedule A (Group A) recorded behavior for six sessions (individual days). Each session was composed of four twenty-minute recording periods. Only one level of the independent variable was encountered within each recording period. Following Schedule B, Group B was employed for four experimental sessions, but all six levels of observer load were encountered in six twenty-minute recording periods. In order to reduce those effects particular to any recording day during any one experimental session the observers recorded each behavior an equal number of periods. Group A had the opportunity to record each behavior during two periods while Group B was expected to record any occurrences of a particular behavior during three different periods of each experimental session.

Previous to each recording period the experimenter would tell which categories were to be recorded by placing a paper with the symbol of each behavior above the
television screen in full view of the observers. The videotape machine was started and the experimenter gave the signal for the observers to begin. The experimenter remained in the area to insure proper operation of the equipment and to guard against collaboration between observers. During each session the observers were allowed one five-minute break, but otherwise the periods followed rapidly one after another excepting time to recycle the equipment. Each videotape used during the sessions was viewed only once by each pair of observers. The tapes were then re-recorded and recycled into the schedule.

During the running of the experiment the observers received no feedback as to the accuracy of their recording and any questions concerning this were answered by the experimenter in general terms such as "it's all right" or "you're doing OK." The experimenter computed the four different agreement scores following the sessions and compiled them into data tables.
RESULTS AND DISCUSSION

Mean agreement scores across sessions and behaviors are presented for both groups (pairs of observers) in Table 1. Each of the four measures of agreement is presented in a separate row. No systematic changes in agreement are apparent as the number of behavior categories increased. Both the I-I and UI scores for both pairs of observers are relatively higher in agreement and more stable than either the SI or F scores. Following the analysis of Hawkins and Dotson (1972) this would be indicative of lower overall rates of behavior. Since lower overall rates of observed behavior would reduce considerably the amount of observer load present throughout the experiment, each behavior category was investigated separately to determine to what extent each was affected by increasing observer load. In the interest of brevity only a representative sample will follow.

The inconsistent results presented in Figures 1 and 2 are representative of much of the data for this experiment. Although no discernible trend can be found in Group A’s agreement for turning around (Figure 1), the SI and F scores of Group B (Figure 2) show a decreasing pattern with larger observer loads producing less reliable recording on the part of one or both of the observers in
that pair. Only the SI scores of Pair A (Figure 1) showed a similar decreasing agreement line.

The agreement scores for recitation show another interesting pattern. Figure 3 presents a clear curvilinear pattern especially in the SI and F scores. This certainly does not support a hypothesis of decreasing performance with greater observer load, though such trends could be interpretable as a reflection of the combined effects of bored distraction and excessive observation load. It is possible that with only one or four behaviors to record the observers became bored and lax in recording while observing a great many behaviors may produce the expected errors due to the large number of consecutive tasks to perform. Informal observation adds some support to this hypothesis: while observing the one and four levels of the independent variable, the observers were often seen slouching and yawning and rubbing their eyes; but during assumedly higher rates of observer load these behaviors did not occur nearly so often and small groans could be heard when the sixteen behaviors condition were initiated.

---

1This trend would be even clearer if a 100 percent agreement score were assigned to the ten behaviors condition. This lack of data point resulted because no occurrence of the behavior was recorded by either observer. A 100 percent score would be reasonable since there were no disagreements and this same situation is handled routinely by the UI and I-I calculations by assigning the perfect agreement figure. On the other hand, using the logic of SI scores, there were no agreements on occurrences, so there can be no agreement score.
Table 1

Mean Agreement Scores Across Sessions and Behaviors

<table>
<thead>
<tr>
<th>Number of Categories</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-I</td>
<td>.92</td>
<td>.91</td>
<td>.93</td>
<td>.84</td>
<td>.92</td>
<td>.88</td>
</tr>
<tr>
<td>SI</td>
<td>.12</td>
<td>.54</td>
<td>.56</td>
<td>.35</td>
<td>.54</td>
<td>.39</td>
</tr>
<tr>
<td>UI</td>
<td>.97</td>
<td>.83</td>
<td>.92</td>
<td>.82</td>
<td>.88</td>
<td>.86</td>
</tr>
<tr>
<td>F</td>
<td>.46</td>
<td>.78</td>
<td>.62</td>
<td>.51</td>
<td>.74</td>
<td>.57</td>
</tr>
</tbody>
</table>

Group A

<table>
<thead>
<tr>
<th>Agreement Calculations</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>10</th>
<th>13</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-I</td>
<td>.97</td>
<td>.88</td>
<td>.93</td>
<td>.74</td>
<td>.88</td>
<td>.88</td>
</tr>
<tr>
<td>SI</td>
<td>.60</td>
<td>.43</td>
<td>.32</td>
<td>.43</td>
<td>.37</td>
<td>.38</td>
</tr>
<tr>
<td>UI</td>
<td>.95</td>
<td>.85</td>
<td>.89</td>
<td>.83</td>
<td>.84</td>
<td>.82</td>
</tr>
<tr>
<td>F</td>
<td>.80</td>
<td>.65</td>
<td>.27</td>
<td>.53</td>
<td>.49</td>
<td>.63</td>
</tr>
</tbody>
</table>

Group B

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 1. Mean per cent I-I agreement (closed circles, solid line), UI agreement (open triangle, broken line), SI agreement (open circles, solid line), and F agreement (closed square, broken line). Four behaviors, seven behaviors, ten behaviors, thirteen behaviors, and six behaviors recorded concurrently by Group A on Turning Around Behavior.
Turning Around Group A

Percent Agreement

Number of Categories

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 2. Mean per cent I-I agreement (closed circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors recorded concurrently by Group B on Turning Around Behavior.
Turning Around Group B

Per Cent Agreement

Number of Categories

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
In Figure 4 one can see that the strong curvilinear trend of Group A is not replicated by the data of Group B. This lack of replication may be the most significant result of this experiment. Almost any conceivable trend may be seen somewhere in the data, yet no single trend overshadows all the rest. Lack of replication between groups may be attributed to any of the classical possibilities. The groups were not run at the same time, often viewed different samples of behavior, and were scheduled according to different matrices. Nevertheless this lack of replication must raise serious doubt as to the magnitude and reliability of any observer load effect.

Another aspect of the data which is somewhat related to the replication issue is that of differences between patterns for the various calculations of the dependent variable. Figure 5 illustrates the basic problem. Aside from the very close agreement between the I-I and UI calculations, the trends exhibited by the calculations are remarkably at odds. The effect of the chosen independent variable, observer load, open observer agreement is highly dependent upon which measure of observer agreement is employed. Entirely different conclusions would be drawn about the effect of observer load depending on which calculation was selected as the "true" dependent variable. This is not a desirable state of affairs. It not only increases the caution which must precede the stating of
Figure 3. Mean per cent I-I agreement (closed circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group A on Recitation Behavior.
Recitation Group A

Per Cent Agreement

Number of Categories
Figure 4. Mean per cent I-I agreement (closed circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four seven, ten, thirteen, and sixteen behaviors observed concurrently by Group B on Recitation Behavior.
Recitation Group B

Number of Categories

Per Cent Agreement

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
conclusive results of this experiment, but of many other experiments which employed only one of the various types of agreement calculations. It seems possible that an entirely different pattern of agreement could have been reflected by a different calculation. Unless these differences in calculations are reconciled at some future point in time, it may be necessary to report agreement scores for all four calculations. Only in this manner could a complete analysis of the reliability of the data emerge.

When it appeared that the concept of observer load did not have functional value, at least within the overall data of this experiment, a reanalysis of the concept produced the possibility that there were two types of observer load. The first type involved the frequency with which a particular behavior occurred. An example of this would be an observer attempting to record a response which occurred so frequently the observer could not keep up with the precise rate of occurrence. The errors produced by the difference in actual response rate and recorded response rate would in all probability produce lower agreement scores between observers. This type of observer load could occur when only one behavior category was being recorded but the load would be considerably aggravated by increasing the number of categories to be recorded.
Figure 5. Mean per cent I-I agreement (closed circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group A on Not Attending Behavior.
Not Attending Group A

Number of Categories

Per Cent Agreement

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The second type of observer load was the original subject of this experiment. Theoretically, it should be possible to expect human observers to record a larger number of behavior categories than is possible for the human sensory apparatus to coordinate. Similar to the case of frequency overload, category overload would produce observer errors and observer disagreement. Logically, the effect would be increased if one or more of the categories were of high frequency.

Since only category overload was controlled or manipulated in this experiment, the interaction of category overload with the uncontrolled frequency overload might conceivably have produced the contradictory results of the experimental data. To determine what effect category overload had upon the observers rate of scoring (the frequency with which the behaviors were recorded), mean per cent of intervals scored tables were compiled for each level of the independent variable and for each pair of observers (see Table 2). The mean per cent of intervals scored figures were derived by summing the number of intervals scored for both observers for each of the behaviors in each of the six levels of the independent variable. The resulting sum is then divided by the number of observers, the number of intervals, and the number of behaviors in that level. This yields a sub-mean for each of the four periods that the observers endured.
each of the six levels of the independent variable. These sub-means are summed and the total divided by four to yield a grand mean percentage of intervals scored for each level of observer load.

It is conceivable that, as the level of category overload increases, the observers might tend to miss a higher percentage of the available data. This would produce a decline in the mean per cent of intervals scored as the number of behavior categories observed increases. As can be seen in Table 2, this is not substantiated by the data. Except for the one behavior observed condition the observers tended score (record) the available intervals at approximately the same rate regardless of the number of behaviors concurrently observed. This is supported by an examination of mean per cent intervals scored data for the five most frequently observed behaviors, Table 3. Here, also, the mean per cent of intervals scored remains relatively stable across the independent variable. This evidence supports the conclusion that the observers did not become less sensitive recording instruments as the number of categories observed increased.

In order to assess the relationship of frequency overload to percentage of agreement, the five behavior categories with the highest mean percentages of intervals scored were grouped to indicate to some extent the effect
Table 2

Mean Per Cent Intervals Scored

<table>
<thead>
<tr>
<th>Number of Categories</th>
<th>Group A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.050</td>
<td>.298</td>
<td>.274</td>
<td>.606</td>
<td>.751</td>
<td>.569</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Categories</th>
<th>Group B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.090</td>
<td>.252</td>
<td>.216</td>
<td>.216</td>
<td>.226</td>
<td>.260</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Mean Per Cent Intervals Scored

<table>
<thead>
<tr>
<th>Most Frequent Behaviors</th>
<th>Number of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Normative Behavior</td>
<td>.732</td>
</tr>
<tr>
<td>Proximity</td>
<td>.920</td>
</tr>
<tr>
<td>Sitting at Table</td>
<td>.204</td>
</tr>
<tr>
<td>Bad Posture</td>
<td>.475</td>
</tr>
<tr>
<td>Inappropriate Fine Motor</td>
<td>.130</td>
</tr>
</tbody>
</table>
of combining the two types of observer load. If any observer effect existed, it seemed logical that this would be the area of the data in which evidence of that effect would be found. The most frequent behaviors in order of frequency were: Normative Behavior; Proximity; Inappropriate Fine Motor; Bad Posture; and Sitting at Table. These will be presented and discussed in order of their rate of occurrence.

Figures 6 and 7 show agreement scores for Normative Behavior, the most frequently recorded response for both groups. Here, again, one can see evidence of conflicting trends between calculations of agreement. More importantly, however, the I-I and SI data from both groups, though irregular, are suggestive of a slight decrease in agreement as more categories of behavior are observed. The SI and I-I scores are in somewhat closer agreement here than in the case of the lower frequency (or duration) behaviors previously presented. Here, also, the UI scores show more extremes of variation not previously seen. This would follow from the fact that as frequency of behavior increases there are fewer unscored intervals and a disagreement on one interval has greater effect on the reliability score, hence the Unscored Interval score will vary more widely.

Data for Proximity behavior are presented in Figures 8 and 9. These data are even more suggestive of
Figure 6. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangles, broken line), SI agreement (open circles, solid line), and F agreement (squares, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group A for Normative Behavior. Mean per cent intervals scored across all levels of independent variable equals 68.0.
Normative Behavior Group A

Mean Per Cent Intervals Scored = .68
Figure 7. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors recorded concurrently by Group B for Normative Behavior. Mean per cent intervals scored across all levels of independent variable equals 72.0.
Normative Behavior Group B

Mean Per Cent Intervals Scored = .72
declining agreement with increasing categories to record. The SI and I-I scores of Group B, Figure 9, have extremely clear trends of decreasing agreement although the UI and F curves indicate a slight recovery after the ten behaviors level. Though the data for Group A are not so clear as in Group B the general trend for the agreement scores is in the decreasing pattern.

The data on the Inappropriate Fine Motor response replicate that of Proximity (see Figure 10 and 11). The I-I scores of both groups appear to be quite clear; agreement declines slightly but consistently as more behaviors are observed. The SI scores of Group A in Figure 9 lend support and replication by exhibiting a dramatic drop in agreement with increasing observer load.

Agreement score trends for Bad Posture for both groups are less easily characterized than the data for the three previous behaviors can be seen in Figures 11 and 12. The trends do not appear to be in support of a declining agreement interpretation, however.

The agreement score patterns for sitting at Table are presented in Figures 13 and 14. Here also the data are less supportive of a declining agreement interpretation fostered by the data of Normative Behavior, Proximity, and Inappropriate Fine Motor. Three of the four agreement calculations of both groups show only the slightest decrease in agreement as the number of
Figure 8. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen and sixteen behaviors observed concurrently by Group A an Proximity Behavior.
Proximity Group A

Mean Per Cent Scored Intervals = .72
Figure 9. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group B for Proximity Behavior.
Proximity Group B

Mean Per Cent Intervals Scored = .64
Figure 10. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group A for Inappropriate Fine Motor.
Inappropriate Fine Motor Group A

Mean Per Cent Intervals Scored = .28
Figure 11. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line) and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group B on Inappropriate Fine Motor.
Inappropriate Fine Motor Group B

Mean Per Cent Intervals Scored = .21
Figure 12. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven, ten, thirteen, and sixteen behaviors observed concurrently by Group A for Bad Posture Behavior.
Bad Posture Group A

Mean Per Cent Intervals Scored = .24
Figure 13. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for four, seven ten, thirteen, and sixteen behaviors observed concurrently conditions by Group B on Bad Posture Behavior.
Bad Posture Group B

Mean Per Cent Intervals Scored = .30
Figure 14. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for the four, seven, ten, thirteen, and sixteen behaviors observed concurrently conditions by Group A for Sitting at Table.
Sitting at Table Group A

Mean Per Cent Intervals Scored = .29
Figure 15. Mean per cent I-I agreement (solid circle, solid line), UI agreement (triangle, broken line), SI agreement (open circle, solid line), and F agreement (square, broken line) for the four, seven, ten, thirteen, and sixteen behaviors observed concurrently conditions by Group B for the Sitting at Table response.
Sitting at Table Group B

Mean Per Cent Intervals Scored = .18

Number of Categories

Per Cent Agreement
categories observed increased.

In general the data from the five behaviors of highest frequency are midly supportive of the conclusion that, in the case of the highest rate or duration behaviors at least, observers agree less often on the recording of a behavior when there are more behaviors to record. In favor of this conclusion is the replication of agreement trends both between calculations and between pairs of observers especially among the three highest frequency behaviors. Less support for this decreasing agreement trend is visible in the agreement scores of sitting at table and Bad Posture although these were somewhat less frequently observed than Normative Behavior, Proximity, and Inappropriate Fine Motor. Detracting from the declining agreement conclusion is the fact that the observers did not observe the highest rate behaviors less often when there were more behaviors to record. Overall the data seem to warrant a cautious conclusion that the levels of observer load employed in this study were effective in decreasing the agreement scores of only the three behaviors having the highest frequency of occurrence. Apparently only when frequency and category overload are combined does observer load have any meaningful effect.
General Discussion

Given that the highest average per cent of intervals recorded during any one condition was only 28.2, serious questions must be raised as to how well the observer load hypothesis was evaluated. It is clear that the experiment dealt with generally low frequency behaviors. Certainly, this stems from the types of behavior definitions selected. The majority of the definitions were selected from published observational codes. To the extent that this was so, the data of this experiment might be generalized to other studies employing those codes. A note of caution must be introduced into the acceptance of high mean agreement scores for those studies which employ a large number of behaviors. Just as in this experiment, low agreement scores an high frequency behaviors could be buried by averaging with many low frequency behaviors. Less caution seems necessary when only low rate behaviors are employed.

Another question on the testing of the observer load effect is raised by the technology employed in the methodology of this experiment. Whereas it was originally believed that scoring sixteen complex behavior definitions in an active subject would be a very difficult task even for a well-equipped observer, it is certainly possible that the use of automatic recording
devices coupled with a drastically restricted environ-
ment simplified the recording task to the extent of
virtually overcoming any observer load effect.

Probably the most necessary refinement to further
research in this area is the development of a new
agreement calculation—a calculation that will be
independent of the rate of behavior, or at least hold
rate of behavior to a certain limited effect. It may be
worth while to employ correlational coefficients, for
example, as is done in much educational and social
psychological research. If a new, more adequate,
calculation cannot be found it may be necessary for
experimenters using observational codes to report a vast
array of agreement figures to provide adequate assurance
of observer reliability.

Future methodological research of the type reported
here might benefit from directly controlling the actual
rate of behaviors to be recorded. It would be interest-
ing to have a set of behaviors acted out according to
some plan which randomized the time between each instance
of behavior while still controlling the rate and dura-
tion. This would allow an analysis of the actual amount
of behavior is missed by observers under a specified
amount of load. It would also control for the effect
that rate has upon the agreement calculation.
An expansion of the design employed in this experiment to include as many as 30 behavior categories would be highly effective in determining the full extent of the observer load effect. Though difficult to implement logistically, such an experiment would yield a definite pattern of the effects of observer load. The results of such an experiment would be highly useful in limiting the size of observer codes to manageable proportions.
BIBLIOGRAPHY


Madsen, Charles H., Becker, Wesley C., and Thomas, Don R. "Rules, Praise, and Ignoring: Elements of Elementary Classroom Control." Journal of Applied Behavior Analysis, I (Summer, 1968), 139-150.


## APPENDIX A

### Definitions

<table>
<thead>
<tr>
<th>Code</th>
<th>Title</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MO</td>
<td>Mouthing Objects:</td>
<td>Bringing thumb, fingers, pencils, or any object in contact with the mouth.</td>
</tr>
<tr>
<td>BP</td>
<td>Bad Posture:</td>
<td>Using arms for support of the upper part of the body when some part of the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>body other than the feet are in contact with the chair or the floor.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Examples: elbows on table, lying on the floor.</td>
</tr>
<tr>
<td>IP</td>
<td>Initiation toPeer:</td>
<td>Subject talks to, or in some way tries for attention of peer.</td>
</tr>
<tr>
<td>PI</td>
<td>Peer Initiation:</td>
<td>Peer talks to, pokes or in some way tries for attention of subject.</td>
</tr>
<tr>
<td>W</td>
<td>Work:</td>
<td>A child may be engaged in appropriate group activity but not working, eg,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>observe a movie. Work means at desk on Academic Projects.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Must work on teacher-assigned task.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recorded when engaged in reading, writing, arithmetic, basic skills.</td>
</tr>
<tr>
<td>NA</td>
<td>Not Attending:</td>
<td>Subject isn't attending to his work or to a lesson being taught, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>May be looking out the window, watching the observer, or other children,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>drawing when he is supposed to be watching teacher demonstrate arithmetic,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>leaning down to tie his shoe, turning his chair.</td>
</tr>
</tbody>
</table>
ST  Sitting at Table:  Subject is sitting on a chair facing a table which is no more than two feet in front of him. Exclude times when his chest is turned more than 90 degrees away from the table, and sitting on the table.

NH  Nodding Head:

P  Proximity:  Scored any time the subject is within arms reach of another person, whether he is sitting, standing, or lying down.

R  Recitation:  Coded whenever subject recites, answers teachers questions, reads out loud, gives a speech, or performs before the class.

PF  Picking Objects Off the Floor:  Lifting any object from surface of floor etc. Don't count taking things off chairs, tables, etc.

NB  Appropriate Group Behavior or Normative Behavior:  Coded whenever the subject's behavior is task-directed activity appropriate for that time and situation. Included would be listening to the teacher explain a lesson, painting during an art class, singing with others during music, lining up with the rest of the class to go for recess, etc. The observer should take care not to include any behavior that might be more appropriately characterized as recitation.

IFM  Inappropriate Fine Motor:  Manipulating small objects or fingers alone which is not task related, and does not make an audible noise. Examples: playing with tokens, turning a pencil, moving and watching fingers, playing with eye glasses. Exceptions: Any objects which are brought in contact with the mouth during the interval.
Ve  Verbalization:  Carrying on conversations with other children when it is not permitted. Answers teacher without raising hand or without being called on; making comments or calling out remarks when no questions have been asked; calling teachers name to get her attention; crying, screaming, singing, whistling, laughing, coughing, or blowing loudly. These responses may be directed to teacher or children.

TA  Turning Around:  Turning head or head and body to look at another person, showing objects to another child, attending to another child. Must be of 4-sec. duration, or more than 90 degrees using desk as a reference. Not rated unless seated. If this response overlaps two time intervals and cannot be rated in the first because it is less than 4-sec. duration, then rate in the interval in which the end of the response occurs.

Vo  Volunteers:  Coded whenever subject raises his hand or in some other manner indicates a desire to recite or do whatever else the teacher may have asked for, e.g., someone to pick up papers; may be either in a class discussion or in a small group.

PC  Physical Contact:
## APPENDIX B

### Schedule A

<table>
<thead>
<tr>
<th>Session</th>
<th>Period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>16</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>10</td>
<td>7</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>16</td>
<td>1</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>7</td>
<td>10</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Session</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>16</td>
<td>10</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>