8-1973

The Replicability of Behavioral Definitions: Implicit vs Explicit Definitions

Robert W. Dobes
Western Michigan University

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

Part of the Experimental Analysis of Behavior Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/2635

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 REPLICABILITY OF BEHAVIORAL DEFINITIONS: IMPLICIT VS EXPLICIT DEFINITIONS

by

Robert W. Dobes

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, 1973
ACKNOWLEDGEMENTS

I would like to thank Dr. Robert Hawkins for his encouragement and advice during the preparation and writing of this thesis. The constructive criticism of Dr. Jack Michael and the support of Dr. Herman Peine were critical in completing the study. My greatest gratitude, however, must be given to the faculty of Milwood Junior High School who graciously volunteered their classrooms and personal support without which this study would not have been possible.

Robert W. Dobes
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 Zeib Road
Ann Arbor, Michigan 48106

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>INTRODUCTION</td>
<td>ii</td>
</tr>
<tr>
<td>GENERAL METHOD</td>
<td>1</td>
</tr>
<tr>
<td>Subjects</td>
<td>1</td>
</tr>
<tr>
<td>Observer Recruitment and Training</td>
<td>1</td>
</tr>
<tr>
<td>EXPERIMENT I</td>
<td>8</td>
</tr>
<tr>
<td>Subjects</td>
<td>8</td>
</tr>
<tr>
<td>Observers</td>
<td>8</td>
</tr>
<tr>
<td>Procedure</td>
<td>8</td>
</tr>
<tr>
<td>Results</td>
<td>13</td>
</tr>
<tr>
<td>Discussion</td>
<td>27</td>
</tr>
<tr>
<td>EXPERIMENT II</td>
<td>32</td>
</tr>
<tr>
<td>Subjects</td>
<td>33</td>
</tr>
<tr>
<td>Observers</td>
<td>33</td>
</tr>
<tr>
<td>Results</td>
<td>34</td>
</tr>
<tr>
<td>Discussion</td>
<td>35</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>38</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>40</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
# LIST OF TABLES AND FIGURES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
</tr>
</tbody>
</table>
INTRODUCTION

A general canon of science is that the methods used to measure the variables of interest—both dependent and independent—must be replicable. If one scientist cannot obtain the same measurements another scientist obtains, no common meaning of terminology is possible and research results from one laboratory are meaningless in another laboratory. Advancement in the understanding of the phenomena toward which that science directs its efforts then become impossible.

In psychology precise, replicable measurement is often achieved in the laboratory by selecting dependent variables that can readily be measured automatically. Typically in basic research, the particular response selected is usually an arbitrary response such as, a bar-press, a key-peck or a similar operant, selected on the basis of how easily and automatically it can be recorded.

The problem is different in applied research and in much sociological research. Here, the experimenter is often interested in particular kinds of responses because of their significance to society or to individuals. Thus, school phobia, disruptive classroom behavior, anxiety, aggression, are behaviors of pragmatic interest and it may be very difficult to measure these behaviors automatically, especially if they are to be studied in the natural setting. In the natural setting, the use of automatic recording devices characteristically found in the laboratory may be cumbersome or inadequate due to the complexity of the setting and/or the complexity of the behavior.
For example, it would be difficult to record a subject's smiling behavior (Hopkins, 1968) or a child's creative block building (Goetz and Baer, 1971) without, at the same time, interfering with the behavior itself. Hence, the reliable use of human beings to measure behavior of other human beings has become a significant area of psychological technology.

In the past, many research psychologists working with significant human behaviors in the natural environment, have specified and evaluated the behaviors to be studied in interpretive terms. These terms or techniques are particularly promulgated by psychodynamic theory which emphasizes assumed inner causes of all behavior, especially "abnormal" behavior. For example, a scientist interested in the problem of "school phobia" in children may have difficulty devising an objective measure of the problem if he considers the real problem to be one of "an unrealistic self-image" (Leventhal and Sills, 1964) or considers the problem the result of hostile impulses of sadomasochistic school personnel toward school phobias which leads children to re-enact in the school the sadomasochistic relationship alleged to exist between mothers and their children (Jarvis, 1964). However, it is readily possible to define "school phobia" in a manner that allows accurate, objective, quantitative measurement of its strength. For example, Ayllon, Smith and Rogers (1970) defined it simply as the frequency of the child's attending school. This re-definition of school phobia into descriptive units (i.e. the presence or absence of the child at school) enables the behavioral scientist to quantitatively and objectively assess the strength of the behavior. If a response is measured in subjective terms there is no assurance
that any other investigator with quite different expectations, biases and predispositions will obtain the same results.

Ayllon and Azrin (1968) repeatedly emphasize, with their "Dimensions of Behavior Rules", the necessity of describing all behaviors to be studied in specific physical units that require a minimum of interpretation. Although behavioral or physical descriptions afford greater research objectivity than do interpretive descriptions, subjectivity and inference may not be entirely eliminated by the use of such physical description. Therefore, a main problem in defining behavior is establishing criteria in a way that two or more observers can agree on the occurrence or non-occurrence of the response.

Some behaviors may require only a minimum of physical description (e.g. smiling or a verbal response such as a question), but most behaviors require a more elaborate description in order to leave as little as possible to the judgement of the observer(s). For example, Harris, Wolf and Baer (1964) defined crying to discriminate it from whining and screaming. Also the crying had to be "loud enough to be heard at least 50 feet away, and of 5-sec. duration". Similarly, a hitting response would have to be clearly defined so that the observer can discriminate hitting from patting or shoving responses. Hence, it is the researcher's burden to adequately define the responses of interest so that his own observers can accurately record the responses with minimal interpretation or subjectivity (non-explicit criteria) and so that other scientists can use the same definitions with reasonable confidence that they are measuring the same behavior. Few researchers in applied behavior analysis seem to take this
responsibility seriously at present.

Since there are few uniformly accepted operational definitions, applied researchers are generally forced to develop their own definitions or amend the definition used by others in order to deal with the individual characteristics of a situation; however, there is no accepted method for determining whether these definitions are sufficiently complete and objective for easy replicability in other laboratories. Unfortunately many definitions given in the published reports appear quite insufficient for use by other scientists.

In the published literature there are at least three basic techniques researchers have used in dealing with response definitions in which these definitions are assumed to be adequate, yet some definitions may be inadequate in practical application. These techniques are (1) not defining the response, (2) grouping undefined responses of similar topography and (3) recording only competing responses.

Consider, for example, a simple and commonly recorded response in the classroom setting, "out of seat" behavior. Some researchers (Broden, Hall, Dunlap and Clark, 1970; and Kunbany, Weiss and Flaggett, 1971) in dealing with this behavior appear to assume that the response is so readily discriminated and its definition so universal in our culture that no explicit definition of the behavior was needed; consequently, they did not include any definition in their published research. Yet while most researchers have included explicit definitions in their published reports, these definitions often appear to leave many decisions regarding specific definitional criteria to
the observer.

Madsen, Becker and Thomas (1968), Thomas, Becker and Armstrong (1968) and O'Leary, Becker, Evans and Saunders (1968) in an apparent attempt to circumvent the problem of writing explicit, precise definitions of specific behaviors, grouped several responses into large classes of behaviors, thus producing a complex multiple category system. Here out of seat behavior was grouped with the behaviors of "standing up, running, jumping . . . moving chair and/or rocking chair, etc." (P.37), under one category coding, gross motor behavior. The rationale of such a grouping was that "certain behaviors, because of their common topography, could be grouped together" (Thomas, et al, 1968, P.37). Since the above authors believe these to be discrete behaviors of similar topography, the grouping of behaviors under one heading, gross motor, allows an observer to avoid the perhaps difficult discriminating necessary to record these individual responses of similar topography. As used here, the Madsen et al technique of grouping not only broadens the physical dimensions of the implied behavior, but the technique also reflects the experimenter's judgment that defining the individual responses making up this behavior category was not necessary. Hence, if the observer's subjective definitions of each component response are not consistent with those of the experimenter's, the absence of a single good definition may allow the observer to develop his own response criterion; the probability of this occurring is compounded with the addition of more undefined component behaviors to the behavior group.
Wahler (1969) dealt with two response classes (disruptive behavior and study behavior) which were similarly composed of several discrete, but not defined, behaviors which included the out of seat response. This technique differs from that of Madsen, et al., in that one behavior category consisted of responses that were incompatible with those of the other behavior category so that the observers recorded either disruptive behavior or study behavior for every observation interval. Since the two response classes were thus mutually exclusive (competing) and topographically different, and since there were no neutral behaviors, there should be less subjectivity resulting from the lack of adequate behavioral definitions (Bijou, Peterson and Ault, 1968). Nevertheless, the component behaviors that made up each response class are not defined, implying that all readers would have the same implicit definition of these behaviors.

In the first example above, out of seat behavior was not defined. In the second example, the behavior was categorized with responses of similar topography but the component responses were neither defined nor described. In the third example, out of seat behavior was contrasted with a competing behavior; yet these competing responses were not defined. All three of these techniques suggest that the authors assume a universal or common response definition; yet there are researchers who do not assume a universal definition for this response and these researchers include in their published research different but precise response definitions for the "out of seat"
behavior. For example, Ramp, Ulrich and Dulaney (1970) and Wolf, Hanley, King, Laccowicz and Giles (1970) defined "out of seat" behavior such that the response criteria required that the seat portion of the child's body not be in contact with any part of the seat of the child's chair. In contrast, Osborne (1969) defined "out of seat" behavior as "any set of movements which results in a subject attaining an upright position without teacher permission" (p.114). Furthermore, "construction of the one piece chair-desk makes it impossible to assume an upright position within the plain chair-desk. Hence, a subject was literally out of his seat before the response criterion was met." These two precise and explicit response definitions can be used to explicate the apparent subjectivity of an undefined "out of seat" response. For example, the grouped or categorized responses of Madsen, et al, (1968)--hopping, out of seat, standing up, jumping, moving or rocking chair, etc.--are differentially used to meet the specific response criteria of Ramp, et al, (1971) and Osborne (1969). That is, only the standing up response is necessary to meet Osborne's criterion of attaining an upright position, whereas, the responses of standing up, jumping, moving or rocking the chair, etc., may all contribute to the Ramp, et al, criterion that the child's seat and the seat of the chair not be touching. Thus had the observers in these two experiments not had an explicit response definition, which of the "grouped" responses would these observers have used as their criterion?

The fact that researchers do not have common response criteria
for what appears to be a simple response, "out of seat" behavior, suggests that observers also may view the behavior differently. Thus, the absence of a definition or the employment of inadequate definitions may leave the observer free to implicitly amend or re-define a behavior so that the behavior now meets the observer's criterion and not the criterion intended by the researcher. While the high agreement obtained when two observers record the same behavior independently would at first appear reassuring regarding the adequacy of the definition of that behavior, there are at least three pitfalls in relying on such reassurance. First, the present methods of measuring inter-observer agreement are often grossly misleading. A frequency recording technique gives only the amount of agreement over the total number of events observed; it does not indicate whether the two observers were recording the same event at the same time (Bijou, et al, 1968). Although interval recording greatly reduces this problem, it has been suggested (Hawkins and Dotson, 1972) that interval agreement scores are frequently inadequate for assessing either the adequacy of the response definition or the alertness and competence of the observers.

Second, the obtained relation may not reflect the general adequacy of the observational process. Romanczyk, Kent, Diament and O'Leary (1971) created an artificial difference between rating capabilities of two reliability assessors. The vocalization and noise definitions of a multiple behavior code were altered for the two assessors such that assessor I was told to rate the category noise as if he had extremely sensitive hearing and assessor II was
told to rate noise as if he could hear only the louder noises that occurred. When an independent observer was introduced, he adjusted his data when recording with assessor I such that the independent observer now appeared to have sensitive hearing. When the independent observer recorded data with assessor II, he adjusted his ratings to correspond to the poor auditory capabilities of that assessor. Furthermore, the authors noted that overall reliability measures were significantly lowered during sessions in which the observers were told that their reliability would not be assessed. This decrease is consistent with the findings of Reid (1970) who reported that observers dropped 25 percentage points in inter-observer agreement from the end of training and overt reliability assessment to the very first day of covert assessment; median reliability remained below criterion during subsequent sessions when the observers were told they would not be assessed. Also, in the genre on confounding observational relationships, Barber and Silver (1970), Kass and O'Leary (1970) and Rosenthal (1966) have demonstrated that observers may adjust their rating to coincide with the expected goals of the researcher. Unfortunately, there is a paucity of all aforementioned methodological research in applied behavior analysis.

Third, the methods by which observers are typically trained may make it possible for them to develop implicit definitions upon which the observers agree very closely (thus the high inter-observer agreement) but that bear only a crude resemblance to the explicit definition given by the experimenter and later presented in the published report of the research. Not only does this make our
scientific precision somewhat illusory, but it may greatly increase the degree to which observers can be biased by other confounding observational variables. Therefore, when humans are observing and recording the behavior under study, any change may represent a change only in their observing and recording responses, rather than in the subject's behavior. "Explicit measurement of the reliability of human observers thus become not merely good technique, but a prime criterion of whether the study was appropriately behavioral" (Baer, Wolf and Risley, 1968, P.93). It is this third limitation of reported reliability scores that the present study was designed to investigate—the problem of implicit versus explicit definitions. If high inter-observer agreement scores reflect only the fact that the two observers have the same implicit definition, the explicit definition reported by the experimenter will be of little value to the general scientific community, for it may only remotely resemble the definition used in obtaining the reported data. In addition, it is likely that observers using inadequate definitions can more readily bias experimental results.

Two experimental designs were employed to investigate the degree to which selected written definitions were able to produce inter-observer agreement without additional elaboration beyond their published form. Definitions were selected from studies that were published in prominent journals. All were from applied behavior analysis in education.

The first design used two independent observer-pairs, which constituted an observer-set. Each independent pair was simultaneously observing the same child and was recording data on the same
behavior. The paradigm attempts to discover whether some published definitions may be sufficiently lacking in precision that observers are forced to make a significant number of decisions regarding definitional criteria, thus developing their own implicit definition.

The second design used three pairs of observers. Each observer-pair observed a different child and recorded data on a different behavior. The paradigm evaluated observer agreement by employing three phases of increasing reliability feedback: (A) when each observer was independent, (B) when the observers were independent, but were allowed feedback on their agreements and (C) when the observers were allowed agreement feedback and were allowed to have collaboration sessions alternate with independent sessions.
GENERAL METHOD

Subjects

The subjects consisted of fifteen children from two eighth-grade and two ninth-grade public junior high school classrooms.

Observer Recruitment and Training

Four observers were contracted through a senior division psychology course in which participating as a research observer was a course requirement for which they received a grade. Ten other observers were recruited independently from the psychology department; they received two hours of academic credit for their participation.

Prior to any training, each observer signed the following pledge:

"I swear that I will observe and record as accurately as I can and that I will follow all procedures requested by __________ even though I do not know the purpose. This includes (a) not comparing my data with my partner or any other observer--either visually or orally--when requested, (b) observing independently when requested (c) not changing my data to make it more like my partner or (d) anything else requested."

Then the observers received a minimum twenty minute introduction on the use of interval recording (Hall, 1971) which included a simple exercise in recording by this technique. All data recording was in blocks of five 10-sec. intervals with a 10-sec. rest interval separating each block. If a response appeared to occur on the border of two 10-sec. intervals, the observers recorded it in the latter interval. After the observers were permanently paired, they recorded three specific and commonly observed behaviors (See Table I) of children in the classroom. While writing this training behavior
code, the experimenter gave particular attention to the development of clear and complete definitions.

Each set of observers sat in a row which, in most instances, was parallel to the wall that faced the children. They chose, from their respective classrooms, three children who could be easily viewed; the same children were used throughout the experiment. The use of three children, rather than one child, minimized the probability of losing a data session due to the absence of a child; also the use of three children minimized the probability of a paucity of data due to a child's low response rate. Child number one was observed during session one; child two was observed during session two and child three was observed during session three. On session four the observers returned to child one and the cycle was repeated. If a child was absent the observers recorded the behavior of the next child in the sequence. The observers always agreed on the child to be observed prior to each session. All observations lasted fifty minutes. In the first three sessions each observer recorded data independently of the other observers. Then the four observers in each set were divided into pairs and, beginning with session four, each pair alternated daily between dependent and independent data collection. During dependent sessions each pair of observers quietly collaborated and jointly recorded the defined behaviors. Collaboration during these sessions allowed observers to quietly discuss with their partner any problems in understanding a response definition or any problems in applying a definition to a questionable behavior observed. If at the end of the session any unresolved
problems were brought to the attention of the experimenter, the experimenter would ask each observer in the pair what he thought would be the best appraisal of or solution to the problem. After both views were expressed the experimenter encouraged the observers to quickly agree upon a solution. When there was more than one pair of observers in a classroom, a cardboard partition was placed between these pairs such that an observer could not communicate with an observer of the other pair.

During independent or "probe" sessions each observer individually recorded the responses. To assure independent observing, cardboard partitions were placed between all observers. Observers were carefully instructed not to discuss their observations or data with any other observer during or after these sessions. Also, no assistance with definition problems was available from the experimenter. After independent sessions the observers tallied their errors. The experimenter then computed agreement scores and informed each pair of their score prior to the beginning of the next session (school day).

All four observers in a classroom synchronized their observations by viewing the same timing mechanism, a large General Electric kitchen clock, and by having one observer always designate a common starting point. The clock was placed such that it was in front of the observers when they were facing the child being observed. To further synchronize the time intervals for the observers, one observer audibly tapped the partition placed between the pairs at the beginning of each 10-sec. rest interval.
Table I

Training definitions

Coding definitions for child behaviors

Writing - anytime that the subject, holding a writing utensil (pen, pencil, crayon), makes contact between the writing end of the utensil and writing material (paper, notebook, workbook, etc).

Exclude writing on the desk and erasing.

Include writing on a book or something of that nature even though you think it inappropriate.

Talk out - any audible vocalization directed at the teacher without the child first raising his hand, or any other audible vocalization directed at other students without permission from the teacher.

Exclude nonsense noises and animal noises etc.

Include instances when the child raises his hand and vocalizes at the same time.

Coding definition for teacher behavior

Talking - any oral sounds involving the vocal cords.

Exclude, especially, coughing, belching, sneezing, and clearing the throat even though they often involve the vocal cords. It is not necessary to be able to understand the words, merely to hear the sound of the subject's voice. Whispering is excluded by the above definition, since it does not involve the vocal cords.

Include simple sounds like "oh", "huh?", "hu", and laughing aloud.
In order to assure that inter-observer agreement scores between the observers would reflect the adequacy of the definitions and not the conscientiousness of the observers, a supplementary observer was employed in each classroom every independent session to record, by a one minute time sampling technique, the attending behavior of the other, primary observers. The definition of this behavior was as follows:

When the observer is looking at the subject, glancing at the clock or writing on his data sheet.

Non-attending: Looking away from the subject, looking at the clock for more than 10-sec. Looking at the data sheet but not writing.

Exclude looking around during the 10-sec. rest interval which is determined when the observer taps on the partition.

Each primary observer was told that the course grade for his observation would be determined by the ratio of attending to non-attending data recorded by the supplementary observer. The agreement scores for attending behavior were computed every other session by a second, independent, supplementary observer. The agreement of these two observers provided a measure of their attentiveness. The primary observers were not told the definition for attending or the recording technique used. After each session the experimenter computed the percent of time devoted to attending by each observer. If overall attending was above 80 percent, the experimenter told each observer his score and complimented him on his performance. Throughout the experiment attending never dropped below 80 percent for any observer.

Reliability

Two techniques were used for assessing agreement between the
primary observers. Traditionally, when interval recording is employed, agreement is computed by dividing the number of agreements by the number of agreements plus disagreements (Hall, Cristler, Chranston, and Tucker, 1970). Here, agreements typically may be either an interval when two observers record the same behavior as occurring or an interval when neither observer records the behavior as occurring. Unfortunately, this technique may afford misleading reliability scores (Hawkins and Dotson, 1972; Bijou, Peterson, Harris, Allen and Johnston, 1969). When a response rate is low, the number of agreements due to the absence of a behavior (a non-response agreement) may yield a very high reliability score even though the observers consistently disagreed upon the few instances when the behavior did occur. Since the present research requires a sensitive index of the adequacy of response definitions, inter-observer agreement was also assessed by a scored-interval technique (Hawkins and Dotson, 1972). The scored-interval method eliminates non-response agreements; thus intervals in which observers did not record a response are not used at data, and the total number of intervals typically used to compute reliability is then reduced. Scored interval reliability is computed by dividing agreements (excluding non-response agreements) by agreements (excluding non-response agreements) plus disagreements. This formula may also be stated as follows: agreements on response occurrence divided by agreements on response occurrence plus disagreements. Since the scored-interval assessment technique is not widely published, a meaningful criterion score was difficult to establish; therefore, an
arbitrarily chosen scored-interval reliability of 50 percent for three consecutive sessions plus an interval-by-interval score of 85 percent for three consecutive sessions was established as an indication that observers were adequately trained. Once this criterion was met by all observers, they received the behavior definitions from the published literature.
EXPERIMENT I

THE DEVELOPMENT OF IMPLICIT DEFINITIONS:
THE EFFECT OF OBSERVER COLLABORATION

Subjects

The subjects consisted of three eight-grade children and three ninth-grade children.

Observers

The research design required a set of two pairs of observers, A-B and C-D. The design was replicated with a second observer-set comprised of four other observers (A₁-B₁ and C₁-D₁). Each set of observers sat in a different classroom with all four observers sitting side-by-side facing the children.

Procedure

The procedure was the same as in observer training but now the observers used response definitions from a published behavior code (Madsen, 1968) consisting of 38 responses (Table II). Although only three response definitions—turning around, appropriate behavior, and academic recognition—were used from the list, the observers were given the entire list each day so that the definitions tested would remain in context with the complete code.

As in training, the observers alternated dependent and independent observation sessions. To assure independent observing, cardboard partitions were placed between observers, but during dependent or collaboration sessions a partition was placed only between the two pairs of observers, such that independence was achieved between pairs but not between the observers within each pair.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
TABLE II
Behavioral Coding Categories for Children

I. Inappropriate Behaviors

A. Gross Motor. Getting out of seat, standing up, running, hopping, skipping, jumping, walking around, moving chair, etc.

B. Object Noise. Tapping pencil or other objects, clapping, tapping feet, rattling or tearing paper, throwing book on desk, slamming desk. Be conservative, only rate if you can hear the noise when eyes are closed. Do not include accidental dropping of objects.

C. Disturbance of Other's Property. Grabbing objects or work, knocking neighbor's books off desk, destroying another's property, pushing with desk (only rate if someone is there). Throwing objects at another person without hitting them.

D. Contact (High and low intensity). Hitting, kicking, shoving, pinching, slapping, striking with object, throwing object which hits another person, poking with objects, biting, pulling hair, touching, patting, etc. Any physical contact is rated.

E. Verbalization. Carrying on conversations with other children when it is not permitted. Answers teacher without being called on; making comments or calling our remarks when no questions have been asked; calling teacher's name to get her attention; crying, screaming, singing, whistling, laughing, coughing, or blowing loudly. These responses may be directed to teacher or children.

F. 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.

G. Other Inappropriate Behavior. Ignores teacher's question or command. Does something different from that directed to do including minor motor behavior such as playing with pencil or eraser when supposed to be writing, coloring while the record is on, doing spelling during the arithmetic lesson, playing with objects. The child involves himself in a task that is not appropriate. Not rated when other inappropriate
TABLE II

G. behaviors are rated. Must be time of task.

H. Mouthing Objects. Bringing thumb, fingers, pencils, or any object in contact with the mouth.

II. Appropriate Behavior. Time on task, e.g., answers question, listens, raises hand, works on assignment. Must include whole 10-sec. interval except for "F" responses of less than 4-sec. duration.

Coding Definitions for Teacher Behavior

I. Teacher Approval following Appropriate Child Behavior

A. Contact. Positive physical contact such as embracing, patting, holding arm or hand, sitting on lap.

B. Praise. Verbal comments indicating approval, commendation or achievement. Examples: that's good, you are doing right, you are studying well, I like you, thank you, you make me happy.

C. Facial attention. Smiling at child.

II. Teacher Disapproval following Appropriate Child Behavior

A. Holding the child. Forcibly holding the child, putting child out in the hall, grabbing, hitting, spanking, slapping, shaking the child.

B. Criticism. Critical comments of high or low intensity, yelling, scolding, raising voice. Examples: that's wrong, don't do that, stop talking, did I call on you, you are wasting your time, don't laugh, you know what you are supposed to do.

C. Threats. Consequences mentioned by the teacher to be used at a later time. If ______ then ________ comments.

D. Facial attention. Frowning or grimacing at a child.

III. Teacher Disapproval following Inappropriate Child Behavior
TABLE II

Same codes as under number II.

IV. Teacher Approval following Inappropriate Child Behavior

Same codes as under number I.

V. "Timeout" Procedures

A. The teacher turns out the lights and says nothing.

B. The teacher turns her back and waits for silence.

C. The teacher stops talking and waits for quiet.

D. Keeping in for recess.

E. Sending child to office.

F. Depriving child in the classroom of some privilege.

VI. Academic Recognition

Calling on a child for an answer. Giving "feedback" for academic correctness.
While dependent, each pair learned to record the behavior by working together and discussing any problems encountered. This simulates a common method of observer training, in which a new observer learns from an experienced one by comparing data with him and discussing any differences. It allows the observers to discuss definitional problems as much as they desire, but has the disadvantage of allowing the development of many "definition rules" (inclusion to and exclusion from the response class being recorded) that are not explicit in the written form of the definition. In the present design, these implicit aspects of each observer's definition would tend to increase the agreement between observers trained together but have unpredictable effect on the agreement between observers trained separately; thus within-pair agreement should tend to be higher than across-pair agreement to the extent that implicitness of definition develops.

After each independent recording session every observer's recordings were compared with every other observer's recordings to determine both agreement within each observer-pair (i.e. A-B and C-D) and agreement across observer-pairs (i.e. A-C, A-D and B-D). Thus it was possible to detect whether each observer-pair was developing its own unique, implicit definition of the behavior or whether the two pairs were in close agreement. If the pairs were in close agreement, the within-pair agreement score should approximate the across-pair agreement score and this should tend to support the assumption that the written definition is an adequate one.

After the first set of definitions from the published literature was adequately tested by comparing within-pair agreement and across-
pair agreement, a second set of three responses taken from another published multiple behavior code\(^1\) (Wasil, Senn, Welch, and Cooper, 1969) was tested using the same observer-pair combinations. The three behavior codes tested were sharing and helping; seeking support, assistance and information and inappropriate sharing and helping. The entire behavior code (Table III) was given each observer daily, with the above three behaviors underlined.

**Results**

Only the results from independent recording sessions will be reported, as scores from dependent sessions would be meaningless. The two within-pair agreement scores obtained from each pair of observers were averaged for every independent session to give within-pair reliability. Across-pair reliability for each independent session was computed by averaging the four reliability scores of the non-paired observer combinations of each set. For each session the mean within-pair scores and mean across-pair scores were plotted (Figures 1, 3, 5, and 7).

Training data are presented in two phases. In the first phase (Sessions 1 and 3) each observer recorded the three training responses independently. In these initial sessions within-pair scores and across-pair scores were computed respective to the pair combinations that were established later in the second training phase.

One training response, talk outs, was not presented in the figures pertaining to observer-set 1 due to the fact that this

\(^1\)This is a modified version of the Coping Analysis Schedule for Educational Setting, CASES, (Spaulding, R.L. 1967).
TABLE III
Child Coding System

Desirable

1. Manipulating and Directing Others: Manipulating, commanding or directing others appropriately; enforcing rules.
2. Self-Directed Activity: Working independently, such as reading, writing, or constructing; continuing to work in the absence of immediate supervision.
3. Sharing and Helping: Contributing ideas, interests, materials; helping others, initiating conversation.
4. Seeking Support, Assistance, and Information: Asking teachers or peers for help, support, direction or explanation.
5. Social Interaction: Cooperative behavior, such as talking, studying, or playing with a peer.
6. Following Directions Passively and Submissively: Following requests, answering direct questions, working only with teacher supervision.

Inappropriate

7. Resisting Authority: More than a 10-sec. delay in carrying out teacher's directions.
8. Observing Passively: Watching others work, "Checking on" activities of adults or peers.
9. 3, 4 and 5: These categories have the same definitions as those with corresponding numbers under the "Desirable" heading, but are coded as inappropriate when they occur at other than the appropriate time or place.

Unacceptable

10. Aggressive Behavior: Direct attack on a child or teacher-grabbing, pushing, hitting, pulling, kicking, name-calling, destroying property.
11. Inappropriate Behavior - Getting Behavior: Activities which seem to result in attention from others such as annoying, bothering, belittling, or criticizing others; noise-making or loud talking.
12. Resisting Authority: Physically resisting instructions or directions, for example - saying "I won't do it" and leaving the room.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Positive</td>
<td>Any verbalization which encourages or approves of the behavior of a child e.g., that's good, you are doing fine.</td>
</tr>
<tr>
<td>14. Neutral</td>
<td>Any statement related to academic work which explains, describes, directs, instructs, or sets limits for a child, e.g., complete pages 6 and 7 in your reading book you may go to the library for thirty minutes.</td>
</tr>
<tr>
<td>15. Question</td>
<td>Any interrogative sentence in which the teacher asks for academic information only, e.g., what is the answer to this problem?</td>
</tr>
<tr>
<td>16. Negative</td>
<td>Any statement that disapproves of a child's behavior. It may be defined by negative content or strong emphasis in speaking, e.g., stop that!, don't do that.</td>
</tr>
</tbody>
</table>
response never was observed to occur. In the latter sessions of figures 1, 3, 5 and 7 the within-pair scores and across-pair scores for the three responses of code 1 (Madsen, et al., 1968) were plotted after which the scores for the three responses of code 2 (Wasil, et al., 1969) were plotted.

The presence, in each session, of a difference between within-scores and across-scores, where this difference consisted of the across-scores being lower, will be referred to here as "disparity". This term will be used to indicate that some degree of definition implicitness was present in a session. The magnitude of the difference between within-pair agreement and across-pair agreement is presented as a "disparity score". A disparity score index (Figures 2, 4, 6 and 8) was computed for each behavior by (A) averaging the within-pair scores across sessions, (B) averaging the across-pair scores across sessions and (C) subtraction of the mean across-pair scores from the mean within-pair scores. These scores will be used as an indication of the degree to which implicitness of response definition developed for any particular definition. The data from Phase 1 training was not used in the computation of disparity scores for these definitions because no inter-observer collaboration had yet occurred and any disparity could only be due to chance.

In Figure 1 the within-pair agreement scores and across-pair agreement scores on all responses are presented in the two training phases, these scores tended to covary such that for each session they increased or decreased within about 10 percent agreement of each other.
The data for the three responses of code 1 indicate that across-scores were only slightly lower than within-pair scores for most sessions. A similar disparity was noted for all three responses of code 2.

The disparity scores (Figure 2) for the training definitions suggest little overall disparity. The scores are as follows: writing, 2 percent and talking, 1 percent. The code 1 definitions suggest little development in overall disparity. The scores are as follows: on task, 1 percent; academic recognition, 4 percent and turning, 10 percent. The code 2 definitions also suggest little development in overall disparity. The scores are as follows: seeking support, 4 percent; inappropriate behavior, 8 percent and redirection, 6 percent.

The scored-interval agreement scores of Figure 3 are based on the same raw data as the scores in Figure 1, except all non-response agreements were ignored. For the training definitions the covariation and proximity of within-scores and across-scores appear more pronounced than in Figure 1.

The scored-interval data for two responses of code 1, on task, and academic recognition, indicate a disparity similar to that noted for the same responses in figure 1, but the scored-interval data for the remaining response, turning, demonstrated an increase in disparity over that evidenced for the same data assessed by interval-by-interval reliability (Figure 1). All of the code 2 response data indicated an increase in disparity over what was evidenced for the same response data assessed by the interval-by-interval method.
Fig. 1. Percent agreement for within-pair observers and across-pair observers of observer set 1 determined by interval-by-interval reliability. Observers who were paired and were allowed to collaborate on their observations during alternate data sessions are within-pair observers (circles). Observers whose partner's data were assessed for reliability with the members of another observer pair in the same experimental setting are across-pair observers (triangles). Across-pair observers were never allowed collaboration sessions. A square represents equal within-pair and across-pair reliability. During training all observers were initially unpaired and were recording data independently. Within-pair agreement and across-pair agreement were computed respective to the observer pairs (paired Os) that were established later in training.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Fig. 2. Disparity scores for observer-set 1 computed from data assessed by interval-by interval reliability. The data from Phase 1 training was not used in the computation of disparity scores for these definitions because no inter-observer collaboration had yet occurred and any disparity could only be due to chance.
MEAN PERCENT DISPARITY

BEHAVIORS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Fig. 3. Percent agreement for within-pair observers and across-pair observers of observer-set 1 determined by scored-interval reliability. Observers who were paired and were allowed to collaborate on their observations during alternate data sessions are within-pair observers (circles). Observers whose partners' data were assessed for reliability with members of another observer pair in the same experimental setting are across-pair observers (triangles). Across-pair observers were never allowed collaboration session. A square represents equal within-pair and across-pair reliability. During training all observers were initially unpaired and were recording data independently. Within-pair agreement and across-pair agreement were computed respective to the observer pairs (paired Os) that were established later in training.
PERCENT AGREEMENT

SESSIONS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Fig. 4. Mean percent disparity for observer-set 1 computed from data assessed by scored-interval reliability. The data from Phase 1 was not used in the computation of disparity scores for these definitions because no inter-observer collaboration had yet occurred and any disparity would only be due to chance.
Fig. 5. Percent agreement for within-pair observers and across-pair observers of observer—set 2 determined by interval-by-interval reliability. Observers who were paired and were allowed to collaborate on their observations during alternate data sessions are within-pair observers (circles). Observers whose partners' data were assessed for reliability with the members of another observer pair in the same experimental setting are across-pair observers (triangles). Across-pair observers were never allowed collaboration sessions. A square represents equal within-pair and across-pair reliability. During training all observers were initially unpaired and were recording data independently. Within-pair agreement and across-pair agreement were computed respective to the observer pairs (paired Os) that were established later in training.
PERCENT AGREEMENT

SESSIONS
Fig. 6. Disparity scores for observer-set 2 computed from data assessed by interval-by-interval reliability. The data from Phase 1 was not used in the computation of disparity scores for these definitions because no inter-observer collaboration had yet occurred and any disparity could only be due to chance.
Fig. 7. Percent agreement for within-pair observers of observer-set 2 determined by scored-interval reliability. Observers who were paired and were allowed to collaborate on their observations during alternate data sessions are within-pair observers (circles). Observers whose partners' data were assessed for reliability with the members of another observer pair in the same experimental setting are across-pair observers (triangles). Across-pair observers were never allowed collaboration sessions. A square represents equal within-pair and across-pair reliability. During training all observers were initially unpaired and were recording data independently; within-pair agreement and across-pair agreement were computed respective to the observer pairs (paired Os) that were established later in training.
Fig. 8. Disparity scores for observer-set 2 computed from data assessed by scored-interval reliability. Observers who were paired and were allowed to collaborate on their observations during alternate data sessions are within-pair observers (circles). Observers whose partners' data were assessed for reliability with the members of another observer pair in the same experimental setting are across-pair observers (triangles). Across-pair observers were never allowed collaboration sessions. A square represents equal within-pair and across-pair reliability. During training all observers were initially unpaired and were recording data independently; within-pair and across-pair agreement was computed respective to the observer pairs (paired Os) that were established later in training.
MEAN PERCENT DISPARITY

BEHAVIORS

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The scored-interval disparity scores (Figure 4) for the training definitions, and for most code definitions, increased at least 100 percent over the scores for the same raw data assessed by interval-by-interval reliability. The scores are as follows: for the training responses, writing was 5 percent and talking was 2 percent. For the Code 1 responses, on task was 2 percent and academic recognition was 6 percent, and turning was 20 percent. For the Code 2 responses, seeking support was 13 percent and inappropriate behavior was 18 percent and redirection was 40 percent.

The experimental procedure was replicated with a second set of observers placed in a different classroom. The results are presented in Figures 5, 6, 7 and 8. In Figure 5 the data were assessed by the interval-by-interval method. Here, within-scores and across-scores of the first training phase covary and approximate the same reliability each session. With the exception of the first session of the first two code 1 responses--turning and on task--and the initial and last session of the first code 2 response--seeking support--the reliability scores for all code definitions indicate a consistent disparity.

The disparity scores (Figures 6) for these response definitions are as follows: for the training responses writing was 3 percent, talk outs was 2 percent and talking was 3 percent. For the Code 1 responses turning was 6 percent, on task was 10 percent and academic recognition was 15 percent. For the Code 2 responses seeking support was 5 percent, inappropriate behavior was 20 percent and redirection was 3 percent.
The scored interval data (Figure 7) for this second set of observers also indicated that the within-scores of the first training phase covaried and approximated, each session, the across-scores by 9 percent or less. Across-scores in the second training phase were 12 percent or less below within-scores. The data for both codes 1 and 2 indicate a more pronounced disparity than that evidenced for the same data assessed by the interval-by-interval method in Figure 5.

The scored-interval disparity scores (Figure 8) for most of the definitions increased over that evidenced when the same raw data were evaluated by the interval-by-interval method (Figure 6). The scored-interval disparity scores are as follows: for the training responses writing was 4 percent, talk outs was 1 percent and talking was 3 percent. For the Code 1 responses turning was 23 percent, on task was 17 percent and academic recognition was 23 percent. For the Code 2 responses seeking support was 31 percent, inappropriate behavior was 34 percent and redirection was 32 percent.

Discussion

Dependent observing and recording produced an agreement between the observers who collaborated (within-pairs) that is not present between the observers who did not collaborate (across-pairs). This was initially evidenced during the training phases. During Phase 1 training observers were unpaired and at all times independent. The corresponding within-scores and across-scores suggest only a chance frequency of across-pair scores being either above or below within-pair scores. Once Phase 2 or collaboration, a component of typical observer training, was introduced across-pair scores fell slightly.
below within-pair scores on almost every occasion. This is interpreted as an indication that some definition implicitness developed and that this implicitness can be attributed to limitations in the clarity and completeness of the written definitions. Of course, it is unlikely that any written definition will be adequate for all observers; there will always be some misinterpretation. But the goal of every researcher writing response definitions should be to so word the definition that individual interpretation is minimized. This not only provides the scientific community with a more replicable procedure for response measurement, but it also reduces the chances of observer bias and the false experimental effects they can produce.

The reliability data gathered from observers using definitions from the published literature were reasonably consistent. The first observer-set evidenced disparity in the code 1 definitions and this disparity was more pronounced with a scored-interval reliability assessment than with an interval by-interval reliability assessment. The code 2 definitions for this observer set appeared to be less explicit than the code 1 definitions and again the evidence for implicitness was more pronounced with the scored-interval reliability assessment. The replication of the experimental procedure with a second observer set produced reliability scores that suggest an overall increase in disparity for both codes 1 and 2 relative to the overall disparity found with the first observer-set. Here also, scored-interval reliability was found to be a more sensitive index of the development of implicit definitions.
Interval-by-interval agreement scores were found to be of questionable value in evaluating definitions. In the sessions in which a response rate was consistently low, the interval-by-interval data (Figures 1, 2, 5 and 6) indicated considerably higher agreement both within-pairs and across-pairs than was evidenced for the same raw data analyzed by a scored-interval technique (Figures 3, 4, 7 and 8). For example, in Figure 5 the interval-by-interval data for the first response of code 2 (i.e. seeking support) and the last response of code 2 (i.e. redirection) demonstrated only slight tendencies in disparity, and the range of both within-scores and across-scores generally fell within 90 percent agreement. In Figure 7, scored-interval agreement for the same raw data indicated a pronounced disparity for the code definitions, and agreement scores generally fell below 50 percent with lower scores usually indicating a diminution of recorded responses by an observer-pair. In some sessions of the seeking support response (Figure 7) no scores could be computed due to the behavior never being recorded by any of the observers of the respective observer-set. That is, since scored-interval reliability eliminates agreements due to the absence of a behavior, data from sessions in which a behavior was not recorded cannot be scored. The same data (or absence of data) would be computed as 100 percent within-pair agreement and across-pair agreement when using the interval-by-interval technique of reliability assessment. Hence, the lack of disparity in the interval-by-interval data of the two examples above, and in many other responses tested, was due to numerous non-response agreements that masked disagreements.
between observer-pairs recording low rate behaviors. The fact that lower agreement scores result when non-response agreements are eliminated and when there are low behavior rates was also suggested by Bijou, et al, (1969) and demonstrated by Hawkins and Dotson (1972).

The masking effect of interval-by-interval reliability assessment also decreased the disparity scores (Figures 2, 4, 6 and 8) for each response occurring at a low rate. This was most pronounced in the redirection response of code 2, Figures 2 and 4. Here mean disparity computed by the interval-by-interval technique was 6 percent; this is contrasted with a mean disparity of 40 percent for the same response data computed by the scored-interval technique.

Scored-interval agreement scores also posed problems in interpreting the present data. The variability of the low scores in the scored-interval assessment tended to be a function of the fluctuations within low behavior rates. Thus, the noted fluctuations in scored-interval agreement scores from session to session were mostly due to these rate differences which changed respective to which child was being observed that session. Obviously, more research is needed to determine the relationship between low behavior rates implicit definitions and scored-interval reliability. The writer speculates that a low behavior rate afforded fewer opportunities for observers to reach agreement on what constitutes a response for each child observed (assuming individual differences in the emission of a behavior), and when observers are recording a poorly defined behavior the number of sessions required to reach satisfactory agreement on a particular child may be substantially increased. Since, for
experiment I, it is the relative relationship between within-pair scores and across-pair scores that was of interest, it was not necessary for observers to have stable and high agreement scores. In fact, any variability in scores may be an added indication of a poorly defined behavior.

In summary, the results of experiment I suggest that a considerable degree of implicitness of definitions developed when observer used the selected published response definitions as a guide in recording. It appears unlikely that this phenomenon is limited to the particular observers employed in the present study, especially since precautions were taken to assure that they did not know the purpose of the experiment and that they recorded the behaviors conscientiously. Also, the overall lack of disparity in the training definitions as compared to the general presence of disparity in the code definitions, should suggest to the reader that better response definitions are not only desirable but that a more conscientious effort is needed to assure the scientific community that definitions developed to allow researchers to measure variables of interest are, in fact, replicable in other settings and with different observers.
EXPERIMENT II

IMPICIT DEFINITIONS DEVELOPED INDIVIDUALLY VS IMPLICIT DEFINITIONS DEVELOPED BETWEEN OBSERVERS

Experiment I demonstrated that considerable definition implicitness developed when observers were allowed to collaborate every other session. Since there are few variables in the behavioral sciences that are encountered on an "all-or-none" basis, the established presence of a variable in question affords only a limited understanding of the relationship between that variable and the behavior studied; consequently any variable of interest can be more fully evaluated when the parameters of its effect are explored. Hence, Experiment II is an attempt to discover how the high reliability scores reported in the Madsen, et al., (1968) study might have developed, given the apparent limitations of the explicit definitions employed. In a sense, Experiment II also serves as a replication of that portion of Experiment I that dealt with the Madsen, 1968 definitions. Specifically this paradigm attempts to measure concomitantly (A) the dissipation of the individually developed implicit definitions of two observers of an observer-pair and (B) the development of a common implicit definition between these two observers.
Method

Subjects

The subjects consisted of three ninth grade students.

Observers

A single organism design was employed in which the "single" organism consisted of one pair of observers. The design employed three observer-pairs, each pair located in a different classroom.

Procedure

The procedure consisted of three phases. Phase I began after observer training. Each pair was given a different response definition from the first behavior code (Table II) and was required to observe that response independently. They were not allowed to compare their data after a session, discuss the observation, or even find out how well their data had agreed. This was continued for several sessions until an estimate of their reliability was obtained. Then Phase II, a procedure resembling "typical" observer-training, was added. The observers were required to compare their data after each session, discuss their disagreements, and tally their errors (as in Experiment 1, reliability scores were not given until the beginning of the next session). When an estimate of their reliability under this condition was obtained over several sessions, Phase III, a second component of "typical" observer training was initiated. In Phase III the observers sat together with no partition between them, comparing and discussing their recordings (especially disagreements) while they recorded. On alternate sessions they recorded independently so that continuous measurement of their reliability could be obtained. School was out for the summer before the second behavior code could be evaluated.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The design provides a more detailed component analysis of the degree of implicitness of observers' definitions. If an observer-pair reaches a high level of agreement on a response while receiving no reliability feedback, and without ever discussing the definition or comparing their data sheets such agreement suggests that the explicit definition of the response is probably adequate (at least for that observer-pair). Any individually developed implicit definitions are either minimal or the two observers have, by chance, individually developed the same implicit definition. If the agreement is initially low but increases during the feedback-only phase, then some degree of definition implicitness may exist. If higher levels of agreement cannot be obtained except by allowing collaboration sessions, then the adequacy of the explicit definition may be questionable.

Results

Since the interval-by-interval technique of data computation produced misleading reliability scores in Experiment I, only data computed by the scored-interval technique are presented for evaluation in Experiment II. The data presented are mean agreement scores for each phase (Figure 9) rather than individual data points representing each session of a phase. The data are evaluated in this fashion because stability in agreement scores could not be established in any of the different phases. As in Experiment I, the variability in scores within each phase was due to differences in the rates of the behaviors exhibited by the three children observed (though each child's behavior was fairly stable in rate throughout the experi-
ment). The fact that each succeeding data point represented a
different child, and hence a different response rate, the joining
of data points in the typical graphic fashion produced variability
that was difficult to interpret. Unfortunately, there were not
enough data points for any individual child to warrant an evaluation
of the data respective to each child for each session of each phase;
but since it was found that the rates for a particular child were
reasonably stable in each phase, averaging the data across sessions,
and thus across children, allowed the relation between definition
implicitness and observer agreement to be more interpretable.

Mean agreement for the turning response, which was measured by
observer-pair I increased from 33 percent in the no-feedback phase to
49 percent in the feedback-only phase. When observers were allowed
to collaborate on alternate sessions, mean agreement increased further
to 59 percent.

For the on task response, mean agreement was 55 percent for the
no-feedback phase, 80 percent for the feedback only phase and 93
percent for the feedback-and-collaboration phase.

For the response of academic recognition mean agreement was 9
percent for the no-feedback phase and 20 percent for the feedback-
only phase. Since the school year was ending at the initiation of
the feedback-and-collaboration phase, academic assignments were
sparse or non-existent. Therefore data collection on academic
recognition was not possible during this last phase.

Discussion

Only the on task definition occasioned a mean agreement score of
50 percent or better in the initial or no-feedback phase. Note that in Experiment I the scored interval disparity indexes (Figures 4 and 8) indicated that this response definition showed the lowest percent disparity (2 percent and 17 percent) of any of the code definitions relative to the respective observer-sets. This combination of a high phase I scored interval reliability score and low disparity scores suggests that the on task definition was the most adequate and explicit definition tested.

The definition for the turning response did not occasion a mean agreement score of 50 percent or better until Phase 3, and the scored-interval disparity indexes from Experiment I for this definition was a moderate 20 percent and 23 percent. These reliability and disparity scores suggest that the turning definition is less explicit.

The third response definition, academic recognition, occasioned the lowest overall mean scores for Phase 1 and 2 and occasioned the least gain in agreement with the initiation of Phase 2, only an 11 percent mean increase. There was no Phase 3 for this definition; therefore the effect of observer collaboration could not be determined. The scored interval disparity indexes Experiment I, for this definition were low (4 percent) to moderate (23 percent). It should also be noted that the observer-pair recording academic recognition reported a general lack of academic assignments in this classroom. This may have been partially due to the school year's coming to an end; therefore recording academics which at this time was a low rate behavior may have differentially affected the observers of
Fig. 9. Mean percent agreement computed from data assessed by scored-interval reliability for each phase of Experiment II.
both experiments, producing scores difficult to interpret.

CONCLUSION

The present research suggests several conclusions: First, it is possible to assess the adequacy of behavioral definitions. Second, some of the behavioral definitions published in the scientific literature leave a considerable amount of the defining process to the observers; thus risking an increase in the probability of observer bias. Third, typical observer training methods, involving the collaboration between observers or at least repeated comparisons of their data through reliability calculations at the end of a session, produce inter-observer agreement that gives an illusory impression that the explicit behavioral definition is adequate. Fourth, the implicit definition that develops in typical observer training is probably a result of both comparing data sheets after recording and collaborating during recording.

Although the findings of the present experiments are limited to only a few of the numerous published definitions of behavior, these findings raise serious doubt as to the validity of many studies in applied behavior analysis. Behavior analysts are typically eager to demonstrate their ability to control significant human behavior and eager to persuade others that systematic application of behavioral principles will give them a greater ability to control such behavior, but it is important that this "salesman" role not be confused with the role of "scientist." Science requires reliable fact where salesmanship requires only faith. As a final note it should be pointed
out that the definitions used from both behavior codes, comprised only a fraction of the many behaviors included in the codes. Since the data from all the sub-classes of behavior were combined by Madsen (Code 1) in reporting their data, the kind of implicitness revealed by the present study could be compounded, with the result that there would be room for much observer bias to be reflected in the reported experimental effects.
REFERENCES


Barber, T. and Silver, M. Fact, fiction, and the experimenter bias effect. Psychological Bulletin Monographs, 1968, 70, No. 6, Part II.

Bijou, S.W., Peterson, R.F. and Ault, Marion H. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. Journal of Applied Behavior Analysis, 1968, 1, 175-191.


Hopkins, B.L. The effect of candy and social reinforcement, instruction, reinforcement and schedule learning on the modification and maintenance of smiling. Journal of Applied Behavior Analysis, 1968, 1, 121-129.


