TREATMENT INTEGRITY IN APPLIED BEHAVIOR ANALYSIS WITH CHILDREN

FRANK M. GRESHAM, KRISTIN A. GANSLE, AND GEORGE H. NOELL

UNIVERSITY OF CALIFORNIA—RIVERSIDE

Functional analysis of behavior depends upon accurate measurement of both independent and dependent variables. Quantifiable and controllable operations that demonstrate these functional relationships are necessary for a science of human behavior. Failure to implement independent variables with integrity threatens the internal and external validity of experiments. A review of all applied behavior analysis studies with children as subjects that have been published in the *Journal of Applied Behavior Analysis* between 1980 and 1990 found that approximately 16% of these studies measured the accuracy of independent variable implementation. Two thirds of these studies did not operationally define the components of the independent variable. Specific recommendations for improving the accuracy of independent variable implementation and for defining independent variables are discussed.

DESCRIPTORS: treatment integrity, children, variables, applied behavior analysis

Central to applied behavior analysis is the demonstration of functional relationships between independent variables and dependent variables. Johnston and Pennypacker (1980) indicated that an independent variable must be represented by known physical parameters of an environmental event. Without clear and unambiguous specification of the independent variable, there can be no definitive conclusions regarding a functional analysis of behavior (Johnston & Pennypacker, 1980; Sidman, 1960).

Peterson, Homer, and Wonderlich (1982) suggest that some operant researchers maintain that steady-state responding in a dependent variable implies the stable and accurate application of the independent variable. However, Peterson et al. (1982) argue convincingly that stability in the dependent variable does not necessarily imply the stable application of the independent variable. For example, tolerance to increasingly higher doses of certain drugs (unstable application of the independent variable) will produce similarities in responses (steady-state responding in the dependent variable).

The degree to which an independent variable is implemented as intended has come to be known

as *treatment integrity* (Peterson et al., 1982; Yeaton & Sechrest, 1981). Treatment integrity is necessary, but not sufficient, for the demonstration of functional relationships between experimenter-manipulated independent variables and dependent variables. That is, some independent variables may be implemented with perfect integrity yet show no functional relationship with dependent variables. Other independent variables may be functionally related to dependent variables; however, this functional relationship may be unknown or may be weak because of the poor integrity with which the independent variable was applied.

Failure to ensure that treatments are implemented as planned poses various threats to the internal and external validity of experiments (Moncher & Prinz, 1991). In terms of internal validity, if significant behavior change occurs and if there are no data concerning the implementation of the independent variable, conclusions drawn from the experiment are questionable because one cannot attribute changes in the dependent variable to manipulated changes in the independent variable. Likewise, if significant behavior change does not occur and if the integrity of treatment is not monitored, one cannot distinguish between an ineffective treatment and an effective treatment implemented with poor integrity (Gresham, 1989; Peterson et al., 1982; Yeaton & Sechrest, 1981).

Requests for reprints should be sent to Frank M. Gresham, School of Education, University of California, Riverside, California 92521.

In terms of external validity, poorly defined, inadequately described, and idiosyncratically implemented treatments make replication and evaluation of treatments difficult (Johnston & Pennypacker, 1980; Moncher & Prinz, 1991). Johnston and Pennypacker define a replication as the degree to which *equivalent environmental manipulations* (i.e., independent variables) associated with earlier observations are duplicated. Replications provide information regarding the generality of a functional relationship over a range of conditions (e.g., subjects, settings, experimenters, etc.). Thus, failure to assess the degree to which treatments are implemented as planned compromises the science of building a replicative history.

Peterson et al. (1982) suggest that a "curious double standard" has evolved in the behavioral literature in which operational definitions and measures of reliability are almost always presented when the behaviors serve as dependent variables but are infrequently presented when these same behaviors serve as independent variables.

The purpose of the present study was to evaluate the degree to which studies (with children as subjects) that were published in the *Journal of Applied Behavior Analysis* (JABA) between 1980 and 1990 assessed or monitored the implementation of independent variables. Peterson et al. (1982) reviewed all studies published in JABA between 1968 and 1980 and found that only 20% of the 539 studies reported data on the integrity of interventions. In addition, over 16% of these studies did not provide an operational definition of the independent variable.

The present review includes studies in JABA conducted with children that have been published since the Peterson et al. (1982) review, and uses a more stringent criterion than that of Peterson et al. for determining whether or not an independent variable was operationally defined. We focus on the research conducted with children because of the large number of studies with children published in JABA. Although not directly comparable to the review by Peterson et al., the current review offers an accurate representation of the studies published in JABA from 1980 to 1990.

METHOD

Criteria for Review

Studies were included in the review based on the following criteria: Only experimental studies that assessed the effect of treatment on behavior were included; brief reports were excluded from the review if these reports were less than three pages in length; the populations studied were children, defined as any subjects under 19 years of age; and the date of publication in JABA was between 1980 and 1990.

Coding Procedures

Studies were coded according to four variables: age level, type of dependent variable, operational definition of the independent variable(s), and assessment of treatment integrity. (A copy of the coding scheme and list of all studies in the review are available from the first author upon request.) The two variables relevant to the present study, operational definition and treatment integrity, are described below.

Operational definition of independent variable(s). Studies were coded "yes" or "no" as to whether the treatment had been operationally defined. The criterion for this category was stated to raters as follows: "If you could replicate this treatment with the information provided, the intervention is to be considered operationally defined." That is, the independent variable was considered to be operationally defined only if its specific verbal, physical, temporal, and spatial parameters were described (e.g., "The experimenter established eye contact with the child, stated 'No' contingent upon a defined response, physically guided the child to the time-out area by the arm, and left the child in time-out for 1 min"). This definition differs from that of Peterson et al. (1982). They used a three-category system of yes, no-unnecessary, and no-necessary, thus making a distinction between independent variables requiring and not requiring operational definitions. We did not use this distinction because we believe that operational definitions are always necessary.

If studies referred to more extensive sources (e.g.,

book chapter, book, technical report, or other article) in which an operational definition was more fully described, the treatment was considered to be operationally defined and was included in the yes category.

Integrity assessment. Studies were classified according to their assessment of treatment integrity. Studies were coded into one of three categories: (a) yes, (b) no, and (c) monitored. If a study assessed integrity and reported an index of treatment integrity expressed as a percentage, the study was included in the yes category. For example, if a study stated that a procedure was implemented on 100% of the occasions during a treatment condition, the study was coded as yes.

If a study made no mention of integrity checks and reported no data on treatment integrity, the study was included in the no category. If a study mentioned that treatment integrity was assessed or monitored but provided no numerical index of integrity expressed as a percentage, it was included in the monitored category. Studies using a treatment protocol or manual that described the procedures used were also included in the monitored category if no specific data regarding treatment integrity were presented.

Training of Raters

Three advanced doctoral candidates in psychology independently coded studies published in *JABA* between 1980 and 1990. The raters were familiar with the literature on applied behavior analysis and had read previous reviews of treatment integrity (e.g., Peterson et al., 1982) as well as other empirical studies published in *JABA*.

Each rater coded one third of the studies in the data base. A random sample of 10% of these studies were also rerated by a fourth advanced doctoral candidate in psychology to assess interrater agreement. Each study was coded according to operational definition of treatment (two categories) and integrity assessment (three categories).

The first author trained the three raters in the use of the coding criteria in a series of four 1-hr sessions spread over 4 weeks. Written materials containing definitions of each category were dis-

tributed to each rater and were discussed in the first session. The raters were given the same two articles to code, and the coding results were discussed in the second session. After discussion of problems and ambiguities in coding, the original code was modified. Two additional articles were distributed to raters, and these articles were coded using the modified coding scheme. The application of the modified coding scheme was discussed in the third training session, the code was further modified to clarify ambiguities, and two additional articles were distributed for coding. In the final training session, the results of the modified code were discussed, and each rater was then assigned one third of the data base to be coded.

Interrater Agreement

Interrater agreement for the four categories in the coding scheme was estimated using two statistics: Cohen's kappa (Cohen, 1960) and percentage agreement. Cohen's kappa estimates were 0.61 for operational definition and 0.83 for integrity assessment. Percentage agreement for the two categories was 81% for operational definition and 93% for integrity assessment. Percentage agreement was calculated by dividing the number of agreements by the number of agreements plus disagreements and multiplying by 100%.

RESULTS

Articles Published by Year

A total of 158 studies meeting the criteria for inclusion in this review were published in *JABA* between 1980 and 1990. On average, *JABA* published approximately 14 child-based experimental studies per year during this time. Fewer studies were published in 1980 and 1985 (11 studies), and more studies were published in 1982 (18 studies) and 1986 (20 studies).

Operational Definitions and Treatment Integrity

Of the 158 child studies published in *JABA* between 1980 and 1990, only 34.2% (54 studies) provided an operational definition of the indepen-

Table 1

Number of Articles Containing an Operational Definition of Independent Variable (by Year)

Year	Yes	%	No	%
1980	2	18.2	9	81.8
1981	3	17.6	14	82.4
1982	15	83.3	3	16.7
1983	0	0	16	100
1984	1	6.7	14	93.3
1985	9	82	2	18
1986	9	45	11	45
1987	3	23	10	77
1988	10	83.3	2	16.7
1989	1	10	9	90
1990	1	6.7	14	93.3
Total	54	34.2	104	65.8

Table 2

Number of Studies in Which Treatment Integrity Was Measured and Reported (by Year)

Year	Yes	%	No	%	Monitored	%
1980	2	18.2	8	72.7	1	9.1
1981	3	17.6	13	76.5	1	5.9
1982	2	11.1	15	83.3	1	5.6
1983	6	37.5	8	50	2	12.5
1984	2	13.3	11	73.3	2	13.3
1985	0	0	11	100	0	0
1986	3	15	15	75	2	10
1987	1	7.7	8	61.5	4	30.8
1988	3	25	9	75	0	0
1989	1	10	8	80	1	10
1990	2	13.3	13	86.7	0	0
Total	25	15.8	119	75.3	14	8.9

dent variable. It should be noted, however, that interobserver agreement for this category was relatively low using kappa (0.61) but approached an acceptable level using percentage agreement (81%).

Table 1 shows the relationship between the year in which studies were published and whether or not the independent variables were operationally defined. Studies were more likely to contain operational definitions in 1982, 1985, and 1988.

Only 25 studies (15.8%) systematically measured and reported levels of treatment integrity for independent variables. Levels of treatment integrity based on these 25 studies ranged from 54% to 100%, with a mean of 93.8% ($SD = 11.63\%$). Almost half of these studies reported integrity levels of 100%. Fourteen studies reported monitoring integrity of the independent variable but provided no data regarding the level of integrity. Table 2 shows the distributions of studies by year and integrity status.

DISCUSSION

Most studies of children published in *JABA* between 1980 and 1990 did not report the accuracy with which independent variables were implemented. Only 25 of 158 experimental studies (15.8%) reported the integrity of independent variable implementation. Peterson et al.'s (1982) review of all experimental studies published in *JABA*

between 1968 and 1980 showed that only 20% of the 539 studies reported data on the integrity of independent variables. Peterson et al. suggested that there was no improvement in methodology since the journal was founded in 1968. Although the present review was restricted to studies of children, we believe that this sample is representative of the status of independent variable measurement in studies published in *JABA* between 1980 and 1990. Our findings indicate that Peterson et al.'s call for increased measurement of independent variables has not been heeded.

Some may argue that treatment effects would not occur in the absence of an accurately implemented independent variable. In this sense, changes in the dependent variable are used as a proxy for appropriate changes in the independent variable. Although behavior change may indicate the application of a treatment, it is not clear whether the degree of behavior change is as strong as it might have been had the treatment been applied with high integrity. In other words, we may know that an effect occurred, but we cannot attribute the intensity of the effect to the description of the independent variable manipulation when no evidence of this manipulation is reported.

It is also possible than an independent variable may be improved in ways that are not explicitly stated. For instance, a researcher may state that a child was given a token contingent on the desired

response. In fact, the child may have received the token coupled with a smile, a wink, a nod, and a pat on the back. A second researcher attempting to replicate the study may find that the token procedures failed to produce the same effects as in the original study. Thus, lack of treatment integrity can result in treatments looking better or worse than they would have if applied as described in the original report.

Replication of experimental effects serves as a self-correcting function and takes into account the procedural variations involved in the administration of independent variables. Replication refers to the repetition of the independent variable(s), whereas reproduction refers to the duplication of the effect(s) of the dependent variable.

Replications of systematic variations of an experimental treatment provide evidence for the robustness of a technique. These replications, however, depend largely on precise descriptions of experimental treatments. Exact replications of an original experimental procedure are virtually impossible, but "equivalent environmental manipulations" can be approximated given precise descriptions of the experimental treatment.

The present review shows that two thirds of the studies of children published in *JABA* from 1980 to 1990 did not provide an operational definition of the experimental treatment. Replication of experimental treatments and measurement of treatment integrity depend upon the degree and specificity with which treatments are operationally defined. Thus, there are reciprocal interrelationships among operational definitions of treatments, replications of treatment procedures, and treatment integrity.

It may be that researchers have operationally defined independent variables, but do not include them in the published descriptions. Journal editors and editorial consultants often require shorter descriptions of methods, given the financial realities of publishing and space allotted to each article. Johnston and Pennypacker (1980) argue, however, that incomplete descriptions of procedural details are almost the equivalent of no description at all and compromise the replicative requirements of

systematic science. Based on the current review, we cannot determine whether most researchers used operationally defined independent variables or vague, ill-defined independent variables.

Recommendations for Measurement of Integrity

The paucity of studies assessing or reporting treatment integrity over the past 11 years limits the confidence in the functional relationships reported. Moreover, the fact that two thirds of the studies failed to define the independent variables sets an upper limit on the degree to which the accuracy of implementation of the independent variable can be measured. Based on the current status of integrity measurement and the operational definitions of independent variables, the following recommendations are made to improve confidence in our research findings.

First, investigators should provide clear, unambiguous, and comprehensive operational definitions of all independent variables used in experimental studies. The same standards with which dependent variables are operationally defined and measured should be applied to independent variables. Independent variables can be defined along four dimensions: verbal, physical, spatial, and temporal.

An example of an adequate operational definition of the components of the independent variable can be found in an investigation by Mace, Page, Ivancic, and O'Brien (1986) that compared time-out with and without contingent delay to decrease disruptive behavior. Mace et al. defined the time-out procedure as follows: (a) Immediately following the occurrence of a target behavior (temporal dimension), (b) the therapist said "No, go to time-out" (verbal dimension), (c) led the child by the arm to a prepositioned time-out chair (physical dimension), and (d) seated the child facing the corner (spatial dimension). (e) If the child's buttocks were raised from the time-out chair or if the child's head was turned more than 45° (spatial dimension), the therapist used the least amount of force necessary to guide compliance with the time-out procedure (physical dimension). (f) At the end

of 2 min (temporal dimension), the therapist turned the time-out chair 45° from the corner (physical and spatial dimensions) and walked away (physical dimension). Failure to provide operational definitions along these four dimensions with the specificity used by Mace *et al.* makes replication and external validation of behavior-analytic investigations difficult.

Second, the integrity with which each component of the independent variable is implemented should be measured using an accuracy criterion. Accuracy refers to the extent to which obtained measures approximate or match the "true" state of nature or the objective, topographic features of behavior (Cone, 1981; Johnston & Pennypacker, 1980). As Peterson *et al.* (1982) suggested, determination of the true value of a dependent variable in nature is not possible, because some portion of the variability in a dependent variable may be influenced by fluctuations in functional relationships between unmeasured and unknown environmental events. In contrast, the "true" value of an independent variable in nature can be operationally defined as the value specified by the experimenter. The extent to which implementation of the independent variable matches or approximates this prespecified value corresponds to its accuracy of implementation.

Third, the level of treatment integrity can be obtained by recording the occurrence and nonoccurrence of the implementation of each treatment component and calculating the percentage of treatment components implemented. Levels of treatment integrity can be calculated for each treatment component within and across sessions (component integrity), as can the accuracy of all components' implementation within sessions (session integrity). Given these two estimates of integrity, failure to find significant treatment effects might be functionally related to poor component integrity over sessions, poor session integrity, or both.

Fourth, two practical considerations should be addressed in measuring treatment integrity. These are (a) the number of observation sessions and (b) reactivity of observation. A major issue in assessing the accuracy of implementation of the independent variable is the representativeness of observational

data taken within and across sessions. Representativeness depends upon both the number of observational sessions and the length of each observational session. Foster and Cone (1986) noted that little attention has been given to the issue of the amount of data needed to produce a representative sample of behavior that will serve as dependent variables. Even less attention has been focused on the amount needed to produce representative samples of independent variables. Generally speaking, the same principles that govern sampling of dependent variables should apply to the sampling of independent variable implementation. That is, one should collect representative samples of integrity of treatment implementation across treatment agents, situations, and sessions (Moncher & Prinz, 1991; Peterson *et al.*, 1982).

Another important consideration in assessing treatment integrity is the potential for reactive effects of the observers' presence in natural settings. Treatment agents may change the accuracy with which independent variables are implemented as a function of being observed and thus produce artificially inflated estimates of treatment integrity. In addition, the same problems of biases in estimating interobserver agreement for dependent variables would hold true for independent variables; these include knowledge of reliability checks, observer instructions, and computation procedures (Foster & Cone, 1986; Moncher & Prinz, 1991; Peterson *et al.*, 1982).

Practical resolutions to potential reactivity of observation in treatment integrity assessments are not easily attained. Several procedures, however, may ameliorate reactive effects. First, researchers can observe within and across sessions on a variable-time schedule and "spot check" the implementation of independent variables (Peterson *et al.*, 1982). Second, observers can try to be as unobtrusive as possible in the treatment environment. Third, researchers can simply not communicate the purpose of observation during the experiment or give treatment providers inaccurate information regarding the purpose of observation and debrief them after the experiment.

Finally, applied behavior analysts should fre-

quently calibrate treatment agents to standard protocols to increase the probability of accurate implementation of the independent variable. Peterson et al. (1982) described the phenomenon of "therapist drift" in which treatment agents deviate from specified treatment protocols, thus compromising treatment effectiveness. Calibration of treatment agents to standard protocols and measuring the degree to which treatment agents adhere to them would provide more assurance that independent variables are being implemented as planned.

In summary, few studies of children published in *JABA* over the past 11 years either measured or monitored the integrity with which independent variables were implemented. The majority of these studies failed to define the independent variables that comprised treatment components. Systematic replications of treatment procedures require operational definitions of treatments that are implemented with integrity. Future work in applied behavior analysis should focus on providing clearer operational definitions of treatments and measurements of treatment integrity to facilitate building a replicative history.

REFERENCES

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Journal of Educational and Psychological Measurement*, 20, 37-46.
- Cone, J. (1981). Psychometric considerations. In M. Hersen & A. Bellack (Eds.), *Behavioral assessment: A practical handbook* (pp. 38-70). New York: Pergamon Press.
- Foster, S., & Cone, J. (1986). Design and use of direct observation. In A. Ciminero, K. Calhoun, & H. Adams (Eds.), *Handbook of behavioral assessment* (Vol. 2, pp. 253-324). New York: Wiley Interscience.
- Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*, 18, 37-50.
- Haynes, S., & O'Brien, W. (1990). Functional analysis in behavior therapy. *Clinical Psychology Review*, 10, 649-668.
- Johnston, J., & Pennypacker, H. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Erlbaum.
- Mace, F. C., Page, T., Ivancic, M., & O'Brien, S. (1986). Effectiveness of brief time-out with and without contingent delay: A comparative analysis. *Journal of Applied Behavior Analysis*, 19, 79-86.
- Moncher, F., & Prinz, R. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, 11, 247-266.
- Peterson, L., Homer, A., & Wonderlich, S. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, 15, 477-492.
- Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.
- Yeaton, W., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, 49, 156-167.

Received June 1, 1992

Initial editorial decision August 26, 1992

Revisions received October 19, 1992; December 10, 1992; February 2, 1993

Final acceptance February 2, 1993

Action Editor, F. Charles Mace