JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1997, 30, 313–326 NUMBER 2 (SUMMER 1997)

TOWARD THE DEVELOPMENT OF STRUCTURED CRITERIA FOR INTERPRETATION OF FUNCTIONAL ANALYSIS DATA

LOUIS P. HAGOPIAN, WAYNE W. FISHER, RACHEL H. THOMPSON, AND JAMIE OWEN-DESCHRYVER

KENNEDY KRIEGER INSTITUTE AND
JOHNS HOPKINS UNIVERSITY SCHOOL OF MEDICINE

BRIAN A. IWATA

UNIVERSITY OF FLORIDA

AND

DAVID P. WACKER

UNIVERSITY OF IOWA

Using functional analysis results to prescribe treatments is the preferred method for developing behavioral interventions. Little is known, however, about the reliability and validity of visual inspection for the interpretation of functional analysis data. The purpose of this investigation was to develop a set of structured criteria for visual inspection of multielement functional analyses that, when applied correctly, would increase interrater agreement and agreement with interpretations reached by expert consensus. In Study 1, 3 predoctoral interns interpreted functional analysis graphs, and interrater agreement was low (M = .46). In Study 2, 64 functional analysis graphs were interpreted by a panel of experts, and then a set of structured criteria were developed that yielded interpretive results similar to those of the panel (exact agreement = .94). In Study 3, the 3 predoctoral interns from Study 1 were trained to use the structured criteria, and the mean interrater agreement coefficient increased to .81. The results suggest that (a) the interpretation of functional analysis data may be less reliable than is generally assumed, (b) decision-making rules used by experts in the interpretation of functional analysis data can be operationalized, and (c) individuals can be trained to apply these rules accurately to increase interrater agreement. Potential uses of the criteria are discussed.

DESCRIPTORS: assessment, functional analysis, visual inspection, interrater agreement

The functional analysis method developed by Iwata, Dorsey, Slifer, Bauman, and Richman (1982/1994) is generally recognized as one of the most significant advancements in applied behavior analysis (Neef, 1994). In the time since its introduction, the procedure has been used extensively and has established a new standard for the assessment of severe behavior disorders. In this analysis, individuals are exposed to a series of test conditions and one control condition, usually in accordance with a multielement design. With an enriched environment as the control condition, the effects of contingent social attention, access to tangible items, escape from demands, and the absence of programmed environmental stimulation on destructive behavior are examined.

The authors thank Jean-Marie Marhefka, Jodi K. Dooling-Litfin, Linda A. LeBlanc, and Matthew L. Remick for their assistance with this project. This investigation was supported in part by Grant MCJ249149-02 from the Maternal and Child Health Service of the U.S. Department of Health and Human Services.

Requests for reprints should be sent to Louis P. Hagopian, Neurobehavioral Unit, The Kennedy Krieger Institute, 707 N. Broadway, Baltimore, Maryland 21205.

Data are interpreted by visual inspection, wherein the analyst examines patterns of responding within and across conditions to determine which, if any, of the variables may be responsible for behavioral maintenance. Finally, an intervention is selected based on the interpretation of the results. For example, a pattern of responding that is characterized by higher rates of aberrant behavior in the demand condition relative to the control condition is generally interpreted as indicating that the behavior is maintained by escape from demands. An intervention such as escape extinction (e.g., Iwata, Pace, Kalsher, Cowdery, & Cataldo, 1990) or functional communication training for escape (e.g., Fisher et al., 1993) may be selected based on this interpretation.

Although Iwata et al. (1982/1994) described various interpretations for different patterns of responding exhibited by their participants, formal and objective procedures for the interpretation of the results using visual inspection have yet to be described in the literature. Because the interpretation of functional analysis results guides treatment selection, accurate and reliable interpretation is critical from both a clinical and a conceptual standpoint. This issue may be particularly important in light of recent trends toward applying functional analysis methods in less controlled community settings (e.g., clinics, schools, group homes; Cooper, Wacker, Sasso, Reimers, & Donn, 1990). As practitioners with varying levels of experience in behavior analysis begin to use functional analysis methods, it becomes increasingly important that structured, objective instructions are available to guide the application of these procedures and interpretation of the results. Unfortunately, little attention has been devoted to the development and articulation of objective methods of data interpretation using visual inspection. The subjective nature of data interpretation using visual inspection is not a limitation specific to functional analysis, but is characteristic of the field of applied behavior analysis as a whole.

Despite the subjectivity of visual inspection, some have argued that it is a reliable and highly conservative method of interpretation for single-case designs (Michael, 1974; Parsonson & Baer, 1986). The available data, however, suggest that interrater agreement of interpretation of single-case data using visual inspection is often less than satisfactory. It is difficult to draw conclusions about the interrater agreement of visual inspection and interpretation of multielement functional analyses because the few studies conducted on visual inspection have used AB designs exclusively. Using a panel of judges experienced in applied behavior analysis, Jones, Weinrott, and Vaught (1978) obtained a surprisingly low mean interrater agreement coefficient of .39 (range, .1 to .79). DeProspero and Cohen (1979), also using experienced judges, obtained an average interrater agreement coefficient of .61 (using a Pearson correlation). This finding is difficult to interpret because judges were required to use a 0 to 100 scale to describe the magnitude of effects rather than judging whether an effect was merely present. Park, Marascuilo, and Gaylord-Ross (1990) obtained a mean interrater agreement of .60, but found that all 5 experienced judges agreed on only 27% of the graphs. Interestingly, only about half of these previously published graphs were judged as showing significant effects. In a study with less experienced judges, Ottenbacher (1990) found interrater agreements between pairs of judges ranging from .53 to .74 on graphs with nonobvious effects, but did not report an overall mean interrater agreement across graphs for all judges. In the one study that has examined intrarater agreement (agreement between three successive interpretations of the same graph by the same individual), Knapp (1983) found that intrarater agreement av-

eraged .78 and did not differ across judges with different levels of experience. Unfortunately, interrater agreement was not reported.

The extent to which the results of these studies on the interrater agreement of visual inspection can be generalized to the Iwata et al. (1982/1994) multielement functional analysis or to the field of applied behavior analysis as a whole remains unknown. It is probable that certain methodological features (e.g., the experience and training of judges, the type of data selected, the exclusive use of AB designs presented in isolation of other data) of previous investigations on interrater agreement of visual inspection may limit their generalizability. Further, interpretation of functional analyses using multielement designs may be more difficult than inspection of AB designs (cf. Iwata, Duncan, Zarcone, Lerman, & Shore, 1994; Vollmer, Marcus, Ringdahl, & Roane, 1995). Interpretation of a multielement functional analysis requires visual inspection of the means, variability, and trends of four or five data paths that may overlap with one another, and differs from judging whether a difference exists across phases in an AB design with a single data path in each phase. Thus, the applicability of these findings to the interpretation of functional analysis data may be more limited. Nevertheless, the available data challenge the assumption that visual inspection of graphed single-case data produces reliable and valid interpretations.

Although some researchers have proposed the use of statistical procedures for the analysis of single-case data, visual inspection remains the preferred method of interpretation. The appropriateness of inferential statistics for analyzing single-case data has been challenged on the basis of statistical as well as conceptual arguments (Baer, 1977; Matyas & Greenwood, 1990). The primary statistical concern is that the serial dependency inherent in single-case data violates the assumption of the independence of observations, one of the most fundamental assumptions of inferential statistics (see Jones, Vaught, & Weinrott, 1977; the reader is referred to Huitema, 1985, for a differing opinion on this matter). The primary objection by behavior analysts is that statistical tests of significance do not aid in the determination of whether an effect is clinically meaningful (see Baer, 1977; Parsonson & Baer, 1986).

The purpose of this investigation was to develop a set of structured criteria for visual inspection of multielement functional analyses that, when applied correctly, would increase interrater agreement and yield interpretations that are consistent with those made by behavior analysts with particular expertise in functional analysis. In Study 1, the level of interrater agreement among 3 psychology interns who used visual inspection procedures (without the aid of any structured criteria) to interpret multielement functional analyses was examined. In Study 2, expert interpretations of a set of 64 functional analysis graphs were obtained. These interpretations were then used to develop a set of objective, structured criteria that yielded similar interpretations. Finally, the interns from Study 1 were trained to apply the structured criteria, and their accuracy and level of interrater agreement were assessed.

STUDY 1: INTERRATER AGREEMENT

Method

Participants

Three predoctoral interns in an APA-approved internship participated. All were matriculated in doctoral programs in psychology and were in the process of successfully completing a 6-month rotation of advanced training in applied behavior analysis on an inpatient unit.

Materials

Functional analyses were completed using methods described by Iwata et al. (1982/1994) for 64 individuals treated for severe destructive behavior in either inpatient or outpatient settings. All clients had been diagnosed with mental retardation and displayed aberrant behaviors including self-injury, aggression, and property destruction. Sessions consisted of a control condition (play) and two to four experimental conditions (demand, alone, attention, and tangible). In all cases, the functional analyses contained 10 sessions in each condition, independent of clarity of the data trends. In some cases, more than 10 sessions per condition were conducted for reasons unrelated to this study. In those cases, only the first 10 sessions from each condition were used in the current investigation.

A graph was generated that depicted the results of each functional analysis. The top panel of Figure 1 shows one of these graphs. Each graph measured 15.3 cm high by 21.6 cm wide and was printed on a single page. Of these 64 functional analyses, 26 were randomly selected.

Procedure

The participants were asked to apply the traditional visual inspection procedures they had been trained to use in graduate school and during their internship to interpret 26 functional analysis graphs during two sessions (13 graphs in each session). For each session, the rater was provided with a packet of 13 graphs and was instructed to select from one of 12 interpretations regarding the function or functions of the target behavior: (a) undifferentiated; (b) maintained by attention; (c) maintained by escape from demands; (d) maintained by tangible reinforcement; (e) maintained by automatic reinforcement; (f) maintained by attention and escape; (g) maintained by attention and tangible reinforcement; (h) maintained by tangible reinforcement and escape; (i) maintained by automatic reinforcement and escape; (j) maintained by automatic reinforcement and attention; (k) maintained by automatic and tangible reinforcement; and (l) maintained by attention, tangible reinforcement, and escape. No other information about the cases was provided.

Results and Discussion

Interrater agreement coefficients between all pairs of judges were determined by dividing the number of exact agreements by the number of agreements plus disagreements. An exact agreement was defined as both raters selecting the same function or functions from the 12 alternatives. The mean interrater agreement was .46 (range, .38 to .50) across pairs of raters.

Across these raters, visual inspection was not a reliable method for interpreting functional analysis data. Although the raters were doctoral students who had received advanced training in behavior analysis, they were not experts, and this may have contributed to the low levels of agreement. In addition, they were not provided with any information about the cases. Although these factors may limit the generalizability of the results, the data are consistent with other findings reported in the literature (e.g., Jones et al., 1978; Park et al., 1990). The results of Study 1 also highlight the need for the development and use of more objective judgment aids to assist in the process of interpretation.

STUDY 2: DEVELOPMENT OF THE STRUCTURED CRITERIA

Method

The purpose of Study 2 was to develop a set of structured criteria for visual inspection of multielement functional analysis data that

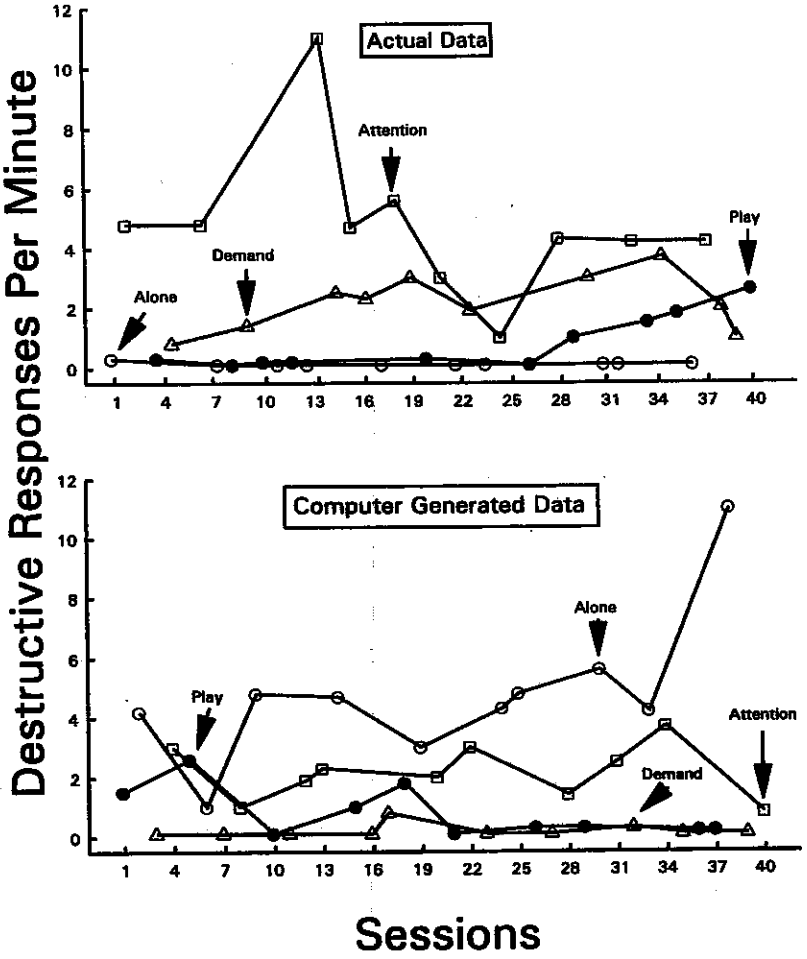

Figure 1. Example of an actual functional analysis graph (top panel) and a computer-generated functional analysis graph (bottom panel) derived from the actual data.

would improve interrater agreement and yield interpretations that were consistent with those made by behavior analysts with particular expertise in functional analysis.

Participants and Materials

The panel of experts consisted of 2 individuals with extensive experience in the use of functional analysis methodology (the fifth and sixth authors). The graphs from the 64 functional analyses described in Study 1 were also used in Study 2.

Development of Structured Criteria

The 2 experts met with the other authors to assist in modifying a preliminary set of criteria for visual inspection and to provide a basis for judging the validity of the criteria. Each of the functional analyses of the 64 cases was visually inspected and discussed. Panelists were asked to make an interpretation of each graph, provide the rationale for their decision, and comment on elements of the data that influenced their interpretation. For each graph, the panel reached consensus regarding the behavioral function or functions.

Agreement between the consensus decision and that derived from the preliminary set of structured criteria was assessed, and disagreements were discussed by the panel.

Most disagreements appeared to be due to the failure of the preliminary criteria to adequately deal with trends, small effect sizes (i.e., small differences between one or more of the test conditions and the control condition), and the interpretation of behavior as maintained by automatic reinforcement. Decision-making rules for these and other situations were incorporated into the final version of the structured criteria, and agreement with the consensus decision was reassessed.

In short, the purpose of the structured criteria was to operationalize the process of visual inspection and interpretation of multielement functional analyses. When using the structured criteria, comparisons are made between each test condition and the play condition (in which there is access to toys, attention, and stimulation, but no demands). The range in which most of the play points lie is defined by drawing upper and lower criterion lines that approximate ± 1 *SD*. The number of points that fall outside this range are then counted. Specific rules are applied for decisions regarding automatic reinforcement, or in cases with trends in the data, low magnitude of effects, or low-rate behavior. The final version of the structured criteria is presented in detail in the Appendix.

Results and Discussion

Agreement between the interpretations reached using the structured criteria (applied by the first author) and the expert consensus was calculated by dividing the number of exact agreements by the number of agreements plus disagreements. The level of exact agreement between interpretations reached using the revised criteria and the consensus interpretations was .94. In the absence of a validated method for interpretation of functional analysis data, expert consensus, which has good face validity, may provide a reasonable standard for comparison. Therefore, these results indicate that the structured criteria have reasonably good concurrent validity with expert consensus.

A few comments about the rules developed for automatic reinforcement should be noted. An interpretation of automatic reinforcement was made under three conditions (see the Appendix). In the third condition (Condition c) for automatic reinforcement, the rates of behavior are relatively high and stable in all conditions. The interpretation of this pattern of responding differs from an interpretation of "undifferentiated" based on the rate and stability of the behavior. Frequent and relatively stable responding, regardless of condition (including the play condition), that occurs over the course of multiple sessions (10 per condition) may suggest that the source of reinforcement is present across time and conditions. When this occurs, automatic reinforcement seems to be a reasonable hypothesis. However, when Condition c is met, additional analyses are recommended to provide further support for this interpretation. For example, Vollmer et al. (1995) have recommended conducting multiple, consecutive alone sessions to help to rule out the possibility that behavior was maintained by social contingencies (e.g., escape from the session room, inadvertent social reinforcement, multiple treatment interference).

Specific rules for dealing with trends in the data, low magnitude of effects, or low-rate behavior were developed and are enumerated in the Appendix. No rules exclusively pertaining to stability of responding were included in the criteria; however, stability is implicitly included. For example, in order for a test condition (e.g., demand) to meet criteria for differentiation (e.g., there are six points above the upper criterion line and one below the lower criterion line), some degree of stability must be present in the data (see the Appendix, General Procedure).

Although these criteria represent an at-

tempt to operationalize and increase the objectivity of visual inspection procedures, the criteria still require some subjective judgments. For example, there is no operational definition for stability, for a "small amount," or for an "overall trend" (see the Appendix).

STUDY 3: TRAINING IN THE STRUCTURED CRITERIA

Method

In Study 3, we assessed the extent to which the 3 predoctoral interns from Study 1, using the structured criteria to interpret functional analysis graphs, would derive interpretations similar to those made by the expert panel (i.e., whether interrater agreement and concurrent validity would increase following training in the application of the structured criteria). The same 3 predoctoral interns who participated in Study 1 also participated in Study 3.

Materials

Two sets of materials were used in Study 3: (a) the graphs from the 64 functional analyses described in Study 1, and (b) a set of 195 computer-generated functional analysis graphs that were similar to the graphs from the original 64 functional analyses. To insure that the computer-generated graphs were similar to actual functional analysis graphs, the data points from the original 64 functional analyses were used to construct the computer-generated graphs. To accomplish this for a particular graph, the order of the data points (i.e., responses per minute from a session) within each condition was randomly reassigned (e.g., the first demand data point might become the seventh demand data point, the fourth might become the third, etc.). This reordering of data points was done with all test conditions (alone, demand, social attention, tangible) and the control condition (play). Next, the rearranged data points from a given test condition (e.g., demand) were randomly assigned to another test condition (e.g., the demand data points might be reassigned as social attention data points). This reassignment of data points to conditions was done for the test conditions (i.e., alone, demand, social attention, tangible), but not for the control condition (play), to eliminate the possibility that the highest rates of destructive behavior might be associated with the play condition (which rarely occurs during actual functional analyses). Thus, after these computerized manipulations were completed, the means and standard deviations of each test condition and the control condition of a computer-generated graph were equivalent to the actual functional analysis graph from which it was generated. Figure 1 shows an example of a computer-generated graph in the bottom panel, which was derived from the actual functional analysis graph depicted in the top panel.

Procedure

Baseline. Each rater was provided with packets of 13 computer-generated graphs and was instructed to select from one of the 12 interpretations (e.g., maintained by attention) regarding behavioral function described in Study 1. No other information about the case was provided. The baseline lengths were staggered across raters in accordance with a multiple baseline across subjects design.

Training in the structured criteria. During training, the first author used didactic instruction, modeling, and practice with feedback to train each rater to apply the structured visual inspection criteria. Training was completed when a participant independently applied the criteria with 100% accuracy to five consecutive graphs. Training times for the 3 participants ranged from approximately 1 to 2 hr. Participants were provided with the criteria (Appendix), an outline of the

Table 1
Outline of Procedures for Applying Structured Criteria for Functional Analyses with 10 Points Per Condition

General procedure
1. Draw upper CL between second and third highest play points.
2. Draw lower CL between second and third lowest play points.
3. Make sure upper CL is at least at 0.5 responses per minute.
4. Count the number of points in each condition that are above the upper CL.
5. Count the number of points in each condition that are below the lower CL (if the lower CL is zero, count zeroes as below the line).
6. For each condition, subtract the number of points that are below the lower CL from the number of points that are above the upper CL. If this number is greater or equal to five for any condition, that condition is considered to be differentiated.

Check for trends for each condition
1. Do at least two of the data points above the upper CL occur in the *second half* of the assessment? If not, there is a downward trend and the condition is not differentiated (apply rules for downward trends).
2. Do all five data points that are above the upper CL occur in the *second half* of the assessment? If so, data points that fall below the lower CL for the first half of the assessment should be ignored and the condition is differentiated (apply rules for upward trends). Also do not adjust upper CL (see exception for low magnitude of effects).

When one or more conditions are differentiated
1. Is alone differentiated along with another condition?
   a. If alone is highest, apply criteria for automatic reinforcement.
   b. If alone is not the highest (relative to other differentiated conditions), apply criteria for multiple maintaining variables.
2. Is there more than one point that is slightly above the upper CL in a condition that meets criteria for differentiation? If so, apply the rules for low magnitude of effects; however, if all five points in the last half are above the upper CL, apply rules for upward trends.
3. Does more than one condition meet criteria for differentiation? If so, apply the rules for multiple maintaining variables, unless
   a. alone is the highest, then apply rules for automatic reinforcement.
   b. alone is the lower of two differentiated conditions, then apply criteria for multiple maintaining variables (i.e., include automatic reinforcement as one of the two functions).
   c. alone is not the highest of three differentiated conditions, then ignore the alone and apply criteria for multiple maintaining variables (i.e., do not include automatic reinforcement as one of the functions).

When no condition is differentiated
1. Are the rates higher in conditions with less stimulation (alone, social attention, and tangible) and lower in demand and play? If so, apply criteria for automatic reinforcement.
2. Are the rates high ($M > 1.5$ per minute) and relatively stable for all conditions, and are there less than five zero points in the whole assessment? If so, apply criteria for automatic reinforcement. Further analysis is also recommended.
3. Is there an overall trend across all conditions without any condition being differentiated? If so, apply the rules for overall trends.
4. Are most of the data points low, with a few high ones? If so, apply the rules for low-rate behavior.

procedures used to apply the criteria (Table 1), and a clear plastic ruler with increments in millimeters.

Posttraining. After the participants had been trained to use the structured criteria, they were presented with packets of 13 of the computer-generated graphs along with the structured criteria. They were instructed to interpret each graph using the structured criteria and to select from one of the 12 alternatives previously listed regarding behavioral function. During posttraining, the participants completed the interpretations independently but received feedback from the first author afterwards. That is, following each posttraining session, the first author reviewed any graphs on which a participant made an interpretation error and provided feedback on the nature of the error. After a participant correctly interpreted 85% of the computer-generated graphs for two consecutive sessions, two additional sessions were conducted using 26 graphs from actual cases (not previously viewed by the participants).

Results and Discussion

The level of agreement between interpretations made by the raters and the interpre-

tations reached using the criteria (applied by the first author) was assessed using exact agreement coefficients as defined in Study 1. The data for each participant are depicted in Figure 2. The first two baseline data points for each participant represent sessions from Study 1, during which data from actual functional analyses were used (13 graphs in each session). The subsequent baseline data points are from interpretations made of the computer-generated graphs. Interpretations made during all sessions were compared with interpretations reached by the first author using the structured criteria. During baseline, the average level of agreement between raters and the first author was .54 (range, .49 to .58 across participants).

Following training, the level of agreement with the structured criteria increased to an average of .90 (range, .86 to .94 across participants). The last two points for each graph in Figure 2 represent sessions in which data from actual cases were used. Mean interrater agreement during the last two sessions, calculated as described in Study 1, increased from .46 in Study 1 to .81 (range, .77 to .85 across pairs of raters).

The results of Study 3 demonstrate that decision-making rules used to interpret multielement functional analysis data can be operationalized and that individuals can be trained to apply these rules with adequate reliability. The high level of agreement (.94) between interpretations reached using the criteria applied by the first author and the expert consensus interpretations suggest that the criteria have reasonable concurrent validity. For each rater, training resulted in immediate and sustained improvements in the level of agreement with interpretations reached using the criteria. In addition, interrater agreement was markedly improved after training (i.e., agreement increased between the participants and the structured criteria as well as among participants).

Despite these positive findings, several limitations of the criteria should be noted. First, the criteria were designed specifically for the interpretation of multielement functional analyses that contain at least 10 points per condition and may not apply to other types of designs or data configurations. Second, interpretations reached using the criteria were based only on the graphed data, which do not account for factors such as intrasession patterns of responding, intensity of responding, or other clinical information. Finally, training and practice are required to apply the criteria correctly.

GENERAL DISCUSSION

The purpose of functional analysis is to identify the variables that are responsible for behavioral maintenance so that an appropriate intervention can be designed. The results of the functional analysis are typically graphed and interpreted by visual inspection of the data. In some cases, because the magnitude of differentiation is great and the data are stable, interpretation is relatively straightforward. In other cases, however, when the differences across conditions are relatively small or the data are variable, interpretation is more difficult. Different individuals may interpret the same data set in different ways and may select different interventions on the basis of their respective interpretations. An erroneous or incomplete interpretation of the functional analysis may lead to an intervention that is either ineffective or, in some instances, iatrogenic (Iwata, Pace, Cowdery, & Miltenberger, 1994). Therefore, accurate and reliable interpretation of functional analysis data is clearly an important step in the assessment process. Despite its importance, none of the published studies on visual inspection have examined the interrater agreement of interpretations of functional analysis data. As noted earlier, the few studies that have been conducted on visual inspection have used sequential designs exclu-

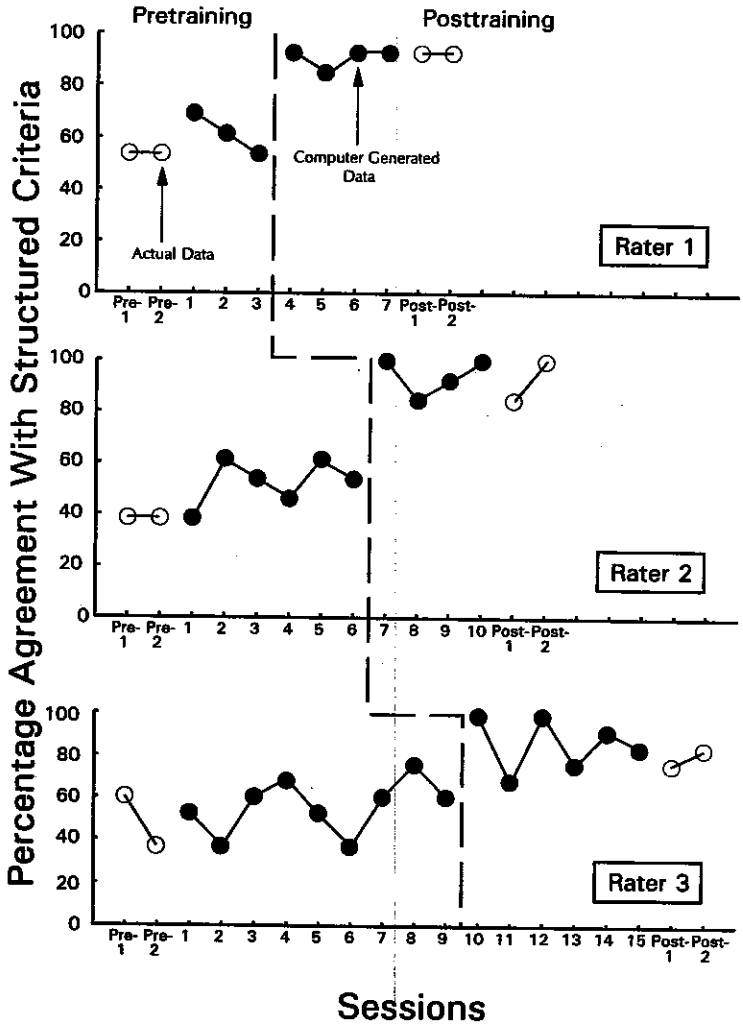

Figure 2. Level of exact agreement with structured criteria across raters for actual functional analysis graphs (open circles) and computer-generated graphs (closed circles).

sively and have reported low to moderate levels of interrater agreement.

The results of Study 1 are generally consistent with these findings and suggest that interpretation of functional analysis data using visual inspection may be less reliable than is typically assumed. These findings should be interpreted with some caution, however, because the raters were not experts and no other clinical information was provided. The results of Study 1, in combination with those reported in the literature, suggest a need for the development and use of more objective judgment aids to guide the process of interpretation of single-case data.

For an objective method of interpretation to be acceptable, it must result in both reliable and valid interpretations. The results of Studies 2 and 3 suggest that (a) decision-making rules used by experts in the interpretation of functional analysis data can be operationalized, (b) interpretations reached using the structured criteria developed in Study 2 have acceptable concurrent validity with expert consensus interpretations, and (c) individuals can be trained to apply these

rules accurately and interrater agreement can be increased to more acceptable levels.

It should be noted, however, that these criteria are not intended to replace visual inspection; rather, they are intended to assist in that process by operationalizing some of the decision-making rules. Any defined set of procedures for interpretation cannot account for all possible situations. In some cases, an interpretation reached using these criteria may be erroneous simply because an important factor is not addressed by the criteria. In addition, exclusive use of the analogue conditions described in the present study or examination of rate data only may not result in valid conclusions about behavioral function. A number of studies have demonstrated that a functional analysis can be conducted using other types of experimental conditions (Bowman, Fisher, Thompson, & Piazza, 1997), experimental designs (Iwata, Duncan, Zarcone, Lerman, & Shore, 1994), and methods of graphically depicting data (Vollmer et al., 1995). Therefore, we strongly caution against relying exclusively on standardized procedures for conducting and interpreting a functional analysis for clinical purposes.

The criteria presented in this study may be most appropriate for training and research purposes. In particular, these criteria may be useful for training behavior analysts and other practitioners to understand and apply the types of decision-making rules that are important to the interpretation of functional analysis data (e.g., using the play condition as the basis for interpreting rates of behavior in other conditions, and considering other factors such as stability, trend, magnitude, and rate).

With the application of functional analysis methods in a broader range of settings with more diverse clinical populations, practitioners with varying levels of expertise in behavior analysis may be conducting functional analyses with increasing frequency. The criteria developed in this investigation may be useful for training such individuals and for guiding their interpretation of functional analysis data. However, rigid and routine application of these or other objective methods for clinical decision-making purposes without the guidance of an appropriately trained behavior analyst is not recommended.

Another potential application of these criteria might include research investigations for which a systematic and uniform method of interpretation may be useful. For example, if two different research centers conducted investigations on operant mechanisms related to the etiology of aggression, the use of structured criteria like those described in the present study could decrease the possibility that inconsistencies in their respective findings could be attributed to differing methods of visual inspection interpretation. Similarly, studies designed to evaluate functional analysis–based treatments might use structured criteria like these to lessen the possibility that treatment failures were due to inaccurate identification of behavioral function.

In sum, although behavior analysts use rigorous data-collection procedures and experimental designs for assessment and treatment evaluation, the procedures used to interpret within-subject data are somewhat subjective. Given that interpretation of functional analysis data is pivotal in the process of treatment selection, we believe that there is a need to examine further the interrater agreement of visual inspection and to work toward the development of procedures to assist in accurate and reliable data interpretation.

REFERENCES

Baer, D. M. (1977). "Perhaps it would be better not to know everything." *Journal of Applied Behavior Analysis, 10,* 167–172.

Bowman, L. G., Fisher, W. W., Thompson, R. H., &

Piazza, C. C. (1997). On the relation of mands and the function of destructive behavior. *Journal of Applied Behavior Analysis, 30,* 251–265.

Cooper, L. J., Wacker, D. P., Sasso, G. M., Reimers, T. M., & Donn, L. K. (1990). Using parents as therapists to evaluate appropriate behavior of their children: Application to a tertiary diagnostic clinic. *Journal of Applied Behavior Analysis, 23,* 285–296.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Fisher, W., Piazza, C., Cataldo, M., Harrell, R., Jefferson, G., & Conner, R. (1993). Functional communication training with and without extinction and punishment. *Journal of Applied Behavior Analysis, 26,* 23–36.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107–118.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis, 27,* 197–209. (Reprinted from *Analysis and Intervention in Developmental Disabilities, 2,* 3–20, 1982)

Iwata, B. A., Duncan, B. A., Zarcone, J. R., Lerman, D. C., & Shore, B. A. (1994). A sequential, test-control methodology for conducting functional analyses of self-injurious behavior. *Behavior Modification, 18,* 289–306.

Iwata, B. A., Pace, G. M., Cowdery, G. E., & Miltenberger, R. G. (1994). What makes extinction work: An analysis of procedural form and function. *Journal of Applied Behavior Analysis, 27,* 131–144.

Iwata, B. A., Pace, G. M., Kalsher, M. J., Cowdery, G. E., & Cataldo, M. F. (1990). Experimental analysis and extinction of self-injurious escape behavior. *Journal of Applied Behavior Analysis, 23,* 11–27.

Jones, R. R., Vaught, R. S., & Weinrott, M. R. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151–156.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277–283.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5,* 155–164.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647–653.

Neef, N. A. (1994). Editor's note. *Journal of Applied Behavior Analysis, 27,* 196.

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: An empirical analysis. *Mental Retardation, 28,* 283–290.

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education, 58,* 311–320.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). New York: Plenum Press.

Vollmer, T. R., Marcus, B. A., Ringdahl, J. E., & Roane, H. S. (1995). Progressing from brief assessments to extended experimental analyses in the evaluation of aberrant behavior. *Journal of Applied Behavior Analysis, 28,* 561–576.

Received August 12, 1996
Initial editorial decision October 15, 1996
Final acceptance January 9, 1997
Action Editor, Timothy R. Vollmer

APPENDIX

Structured Criteria for Visual Inspection of Multielement Functional Analysis Data for Analyses with 10 Points per Condition

General Procedure

An upper criterion line (CL) and a lower CL are drawn to approximately 1 *SD* above and below the mean of the control condition (play). The lines are drawn based on the number of data points that would hypothetically fall beyond 1 *SD,* assuming a normal distribution of the play data points. Thus, the upper CL for 10 points is drawn between the second and third highest points, and the lower CL is drawn between the second and third lowest points. Criterion for differentiation is based on the number of data points for each condition that fall beyond the CLs. Differentiation is said to occur if at least five more data points from a test condition fall above the upper CL than fall below the lower CL. If the lower CL is zero, count each zero point as below the

lower CL. Note that the minimum upper CL is drawn at 0.5 responses per minute.

Rules for Automatic Reinforcement

Score functional analysis as automatic only if (a) alone is the highest condition and is signficantly higher than play; (b) the rates of behavior tend to be higher (across most sessions) in conditions with less external stimulation (alone, social attention, and tangible) and lower in the conditions with higher external stimulation (demand and play); or (c) all conditions are high and relatively stable with no overall trends (the mean of all conditions is greater than or equal to approximately 1.5 per minute), and there are less than five zero points. Note that if Condition c criteria are met, further analysis is recommended.

Rules for Trends

Downward trends. At least two of the data points above the upper CL must occur in the second half of the assessment; otherwise there is a downward trend and the condition is not differentiated. *Exception:* For the demand and tangible conditions, do not apply the differentiation rules for downward trends if there is a decreasing trend to an efficient rate of responding (e.g., if escape or tangible items are provided for 30 s contingent on behavior, an efficient rate of responding would be two per minute).

Upward trends. If all five data points that are above the upper CL occur in the second half of the assessment, this is an upward trend, and data points that fall below the lower CL for the first half of the assessment should be ignored (i.e., the condition is differentiated). Also, the upper CL should not be adjusted in this case (see rules for low magnitude of effects).

Overall trends. If there is an overall trend across most of the conditions (including play), any condition that is consistently higher than play over the course of the assessment meets criterion for differentiation.

Rules for Low-Rate Behavior

In cases in which most of the data points are low, the condition in which all or most of the higher rate behavior occurs is considered to be differentiated (i.e., more than half of the higher rate sessions occur in one condition *and* more than half of the total number of behaviors in the higher rate sessions occur in that same condition). However, one of those high points must occur in the last half of the assessment.

Rules for Low Magnitude of Effects

In cases in which a condition meets criteria for differentiation but more than one of the points are above the upper CL *by only a small amount* (i.e., the magnitude of differentiation is relatively low), raise the upper CL by 20% (for the condition with the low magnitude of effects). Use this adjusted upper CL for determining differentiation for that condition instead of the regular upper CL. *Exception:* If there are five points above the upper CL in the last half of the assessment (i.e., the condition meets criteria for an upward trend), do not apply the differentiation rules for low magnitude (do not adjust the upper CL).

Multiple Maintaining Variables

In cases in which more than one condition meets criteria for differentiation, score the analysis as multiply maintained (unless the highest is alone; then score it *only* as automatic). If there are three differentiated conditions and the alone condition is not the highest among those, ignore the alone condition (e.g., do not score it as automatic, attention, and tangible; score it as attention and tangible). If there are two differentiated conditions and the alone is the lower of the two, score it as both automatic and the other condition.

STUDY QUESTIONS

1. What is the primary advantage and disadvantage of using statistical analysis, rather than visual inspection, for interpreting data from single-subject designs?

2. Why is the interpretation of functional analysis data from multielement designs typically more difficult than it is for data from AB designs?

3. In Study 1, low agreement (M = 46%) was obtained between raters who were asked to render one of 12 interpretations for each of 26 sets of functional analysis data. What simple procedure might have increased interrater agreement?

4. Describe the steps taken in developing the structured criteria for data interpretation.

5. What general strategy for data interpretation is operationalized in the structured criteria?

6. How did the authors develop the computer-generated functional analysis graphs used in Study 3?

7. Briefly summarize the procedures followed and the results obtained in Study 3.

8. For what specific applications did the authors recommend using structured criteria such as those presented in the study?

Questions prepared by Iser DeLeon and SungWoo Kahng, University of Florida