JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1982, 15, 415-421 NUMBER 3 (FALL 1982)

INTERVENTION EFFECTS AND RELATIVE VARIATION AS DIMENSIONS IN EXPERTS' USE OF VISUAL INFERENCE

MICHAEL J. FURLONG AND BRUCE E. WAMPOLD

STATE DEPARTMENT OF EDUCATION, HONOLULU, HAWAII
AND UNIVERSITY OF UTAH

Recent research indicates that when analyzing graphically presented single-subject data, subjects trained in visual inference appear to attend to large changes between phases regardless of relative variation and do not differentiate among common intervention effect patterns. In this follow-up study, experts in applied behavior analysis completed a free-sort task designed to assess the effects of these dimensions on their use of visual inference. The results indicate that they tended to differentiate among common intervention effect patterns but did not attend to relative variation in the data.

DESCRIPTORS: visual inference, data analysis, experimental effects

Although visual inference is the predominant mode of data analysis for single-subject designs (Kratochwill & Brody, 1978), only recently has this procedure been empirically analyzed. The nature of visual inference judgments has been investigated by comparing the conclusions derived from visual inference with those of statistical analysis or by evaluating interjudge agreement. Jones, Weinrott, and Vaught (1978) compared judges' use of visual inference to time-series analysis and found that there was little agreement between these two modes of data analysis and that interjudge agreement was low. DeProspero and Cohen (1979) developed hypothetical graphs, and reviewers of behavioral psychology journals rated the degree of "experimental control" shown in the graphic data; they also found that interjudge agreement was low. The results of these studies indicate that visual inference may be an unreliable mode of data analysis.

Research on the reliability of visual inference, however, has shed little light on how individuals visually analyze time-series data. One investigation of this topic (Wampold & Furlong, 1981) assumed that when a graph is viewed it is compared to the common intervention effect patterns discussed in the single-subject methodological literature (see, for example, Parsonson & Baer, 1978, pp. 123-129) and is classified as an example of the particular pattern (e.g., positive change in level, immediate and lasting effect) best fitting the data. Thus, each graph is conceptualized as a transformation of one of the idealized data patterns. The pattern of changes across phases of a time series and the amount of variation in the data were evaluated in our previous study because they have been emphasized as important elements of visual inference (cf. Hersen & Barlow, 1976; Parsonson & Baer, 1978). Two-phase graph sets were randomly generated from idealized intervention effect patterns and presented to two groups of graduate students; one group had completed a single-subject research seminar and the other a course in multivariate statistics. The students free-sorted a set of graphs into classes that they believed represented similar experimental effects. It was found that the students trained in visual inference attended more to large changes between phases of the graphic data regardless of relative variation than did the students trained in multivariate statistics.

Special appreciation is extended to Professor Lawrence J. Hubert for his assistance during this study. Reprints and a detailed description of the algorithms used to generate the data sets and the computer program are available from Bruce E. Wampold at the Department of Educational Psychology, University of Utah, Salt Lake City, Utah 84112.

In addition, both groups were unable to differentiate among graphs involving a change in level and trend from those showing only a change in trend.

The theoretical model used in our previous study predicts that the ability to classify data patterns should improve as exposure to transformations of the common intervention effect patterns increases (Posner, 1973). Thus, the results reported for students of visual inference may not generalize to experts of single-subject methodology. The purpose of this study was to examine the ability of experts in applied behavior analysis to view graphs, classify them as examples of common intervention effect patterns, and to determine how variation in the data affects these judgments.

METHOD

Participants

Ten members of the *Journal of Applied Behavior Analysis* (*JABA*) editorial board (of 36 who were randomly selected and asked to participate) completed the experimental task. They represented a highly qualified group with nine having a Ph.D. degree and one an M.D. degree. All reviewers appeared to be well versed in single-subject methodology; eight had taken and/or taught a single-subject design course and had published an average of 8.4 studies using single-subject procedures.

Procedure

Intervention effects reflecting a change in level, trend, or both were selected as the idealized data patterns from which the time-series data were generated (see Figure 1). The idealized data patterns had no slope in the baseline (phase A), and all changes were immediate, lasting, constant, and in the upward direction (phase B). To generate the actual graphs presented to the experts, these idealized data patterns were modified by using three transformation procedures. A *standard transformation* merely added a randomly and independently selected normal deviate (i.e., white noise) to each value in the idealized data pattern. A *scaling transformation* was formed from the standard transformation by "stretching" it along the vertical axis. Specifically, the standard deviation of the random deviates and the parameter used to generate the change in level and/or trend was multiplied by a constant. A *variation transformation* was also formed from the standard transformation by multiplying the standard deviation of the random deviates by the same constant, but the parameter used to generate the change in level and/or trend was left unaltered.

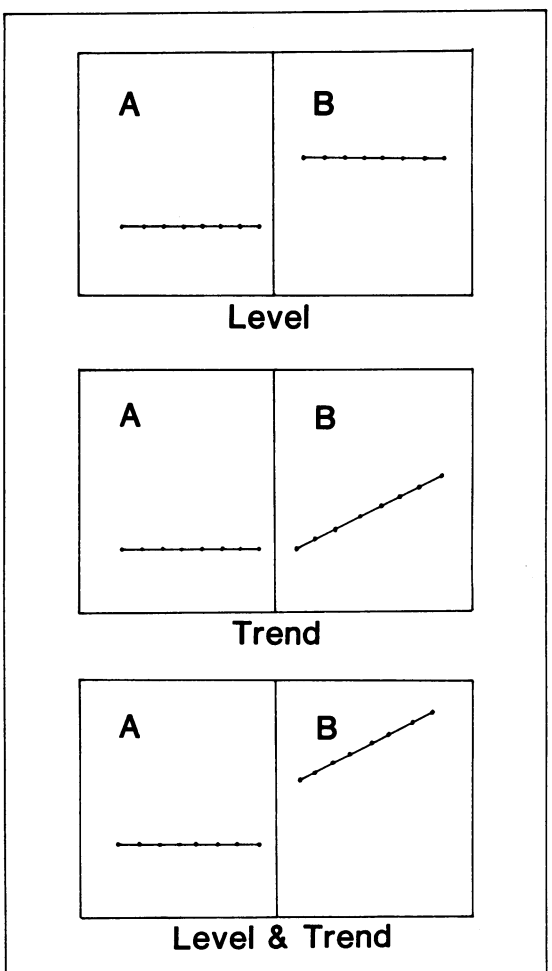

Fig. 1. Change in level, change in trend, and change in level and trend idealized intervention effect patterns (the idealized data set patterns, from which the transformations were developed, were not presented to the participants).

The modifications to the idealized data patterns are best discussed in terms of the three transformation procedures. The standard and scaling transformations were mathematically equivalent because the parameters used to generate them differed only by a multiplicative constant. Thus, the absolute size of the change in level and/or trend of the scaling transformation was larger than that of its associated standard transformation, but, when compared to variability, the relative size of the intervention effect was exactly the same. In contrast, the variation transformation differed from the standard and scaling transformations in that only the variation parameter was increased. Thus, the size of the experimental effect (change in level and/or trend) when compared with variation was smaller in the variation transformation than for the associated standard and scaling transformation.

A computer program was written to select randomly the change in level and/or trend, the variation, the multiplicative (scaling) constant, as well as the number of data points (range = 5 to 10) in each phase for each set of associated standard, scaling, and variation transformations. (Given the nature of the mathematical algorithms used to select the data parameters, it is possible that by chance a small number of the graphs did not reflect the desired patterns. For this reason, each reviewer received a randomly selected subset of the graphs, thus ensuring that this unavoidable factor did not bias the sorting task.) A graph set thus consisted of a standard transformation, a scaling transformation, and a variation transformation, all of which had the same underlying intervention effect parameters. This random data generation procedure was repeated 15 times for each of the idealized data patterns. Thus, 45 graph sets (135 separate graphs) were created. Figure 2 presents an example of a graph set for each of the intervention effect patterns. It is important to note, once again, that within each graph set, the standard and scaling transformations were mathematically equivalent and any differences in the up-and-down profile of the data were due merely to chance. Each data set was drawn on a 5- × 8-inch index card with phase A and phase B sepa-

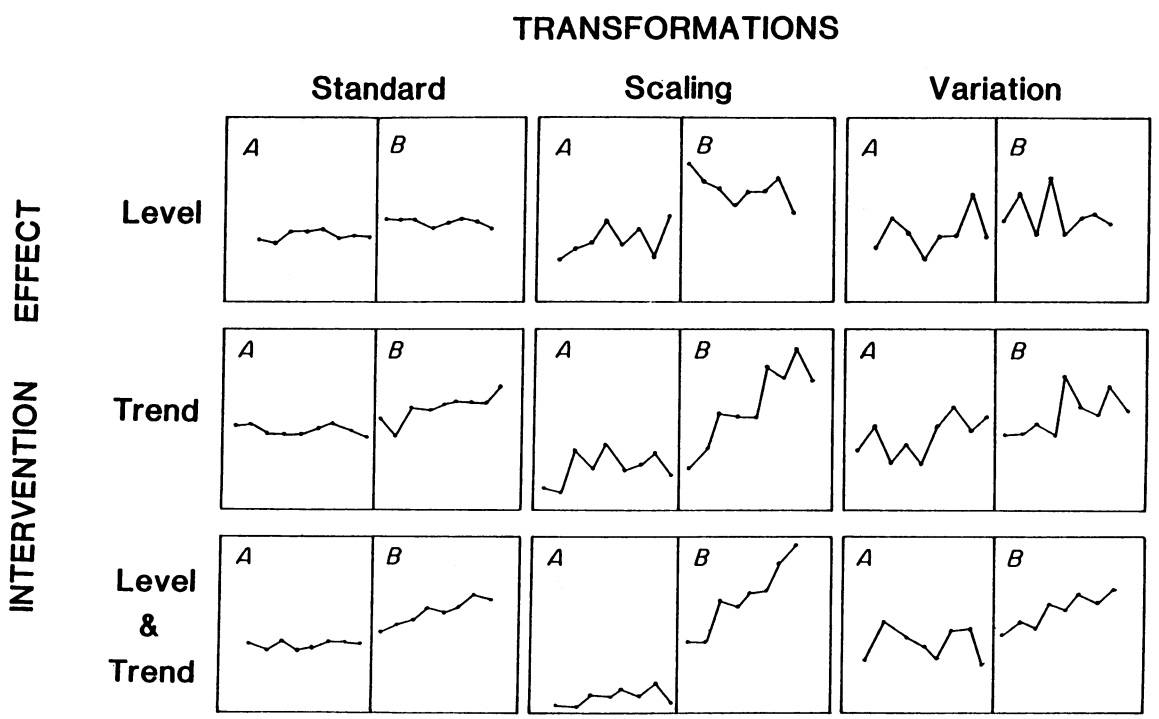

Fig. 2. Examples of graph sets (actual graph size was 5" × 8").

rated by a solid vertical line. The mean of the data values was centered on the card and no abscissa or ordinate was drawn, although the experts were instructed that the edges of the card served this purpose.

The 10 *JABA* reviewers were each mailed 12 randomly selected graph sets, four of each intervention effect pattern, for a total of 36 graphs. They were requested to free-sort the graphs into disjoint classes that demonstrated "similar experimental effects" and were allowed to make as many or as few classes as they wanted. This sorting instruction was used so that the experts would be allowed maximum flexibility to impose their idiosyncratic visual inference strategies on the experimental stimuli. After completing the sorting task, the reviewers bound each pile of graphs with a rubber band and returned them to the experimenters.

RESULTS

Each reviewer's classification of the graphs was summarized in an individual 36×36 symmetric matrix, with each cell entry assigned a value of one if two given graphs were placed in the same class and zero if not. The individual matrices were then compared to three matrix partitions that reflected possible sorting strategies that could have been used by the reviewers.

Strategy 1. If the reviewers used this classification strategy they would have sorted the graphs into two piles, one containing the 24 standard and scaling transformations and another containing the 12 variation transformations. This sorting strategy respects the equivalence of the standard and scaling graphs because the relative changes from phase A to phase B in both types of graphs were exactly the same.

Strategy 2. If the reviewers used this classification strategy they would have sorted the graphs into two piles, one containing the 12 scaling transformations and a second containing the 24 standard and variation transformations. This sorting rationale attends only to the absolute size of effect and not size of effect in relation to variation. A reviewer using this strategy would have been looking for large, dramatic effects.

Strategy 3. If the reviewers used this classification strategy they would have sorted the graphs into three piles defined by the three intervention effect patterns: change in level, trend, or both. This sorting rationale reflects the ability to differentiate among the intervention effect patterns.

Each reviewer's matrix was compared to each of the classification matrices by computing a cross product index that indicates the degree of similarity between the two matrices (Hubert & Levin, 1976). To compare the classification strategies used by the reviewers the index was standardized using formulas for its mean and variance (Hubert & Schultz, 1976). Thus, the standardized index indicates the degree to which each reviewer respected the classification strategies; large values ($\geqq 2.00$) indicate that the strategy was applied, whereas small values indicate that the strategy was ignored (i.e., the reviewer's classification results were random with respect to the strategy).

The proportion of reviewers who obtained standardized index values greater than or equal to 2.00 was obtained for each strategy matrix. For comparative purposes, the proportions of students trained in single-subject research and multivariate statistics who also obtained this criterion level were taken from data reported in our previous study (Wampold & Furlong, 1981). These proportions are shown in Table 1.

The results revealed that 20%, 60%, and 70% of the experts obtained index values of 2.00 or greater for strategies 1, 2, and 3, respectively. It appears, therefore, that as a group, the reviewers did not totally respond to the equivalence of the standard and scaling transformations (Strategy 1) and attended more to the absolute size of the change between phases without taking relative variation into account (Strategy 2). Of the three strategies evaluated, the most reasonable global description of the reviewers' classification rationale, however, was that they attended

Table 1

Proportion of participants obtaining standardized index values exceeding 2.00 for each strategy.

	Group		
Strategy	n = 1 *Trained*[a] (N = 14)	*Multivariate Trained*[a] (N = 10)	JABA *Reviewers* (N = 10)
1	.07	.40	.20
2	.64	.20	.60
3	.36	.60	.70

[a]Data reported in Wampold & Furlong (1981).

primarily to the intervention effect patterns (Strategy 3).

To enable the reader to develop a more complete picture of the experts' performance on the experimental task, the standardized index values for each expert are presented in Table 2. These data corroborate the previous data; that is, most of the experts were able to identify the intervention effect patterns (Strategy 3), but attended to the absolute size of the effect (Strategy 2) more than the relationship between the size of the effect and variation (Strategy 1).

A secondary analysis revealed that the Pearson product moment correlations among the index values were as follows: Strategies 1 and 2, $r = -.64$ ($p < .05$); Strategies 1 and 3, $r = .42$ (n.s.); Strategies 2 and 3, $r = -.77$ ($p < .01$).

Table 2

Expert's Standardized Index Values for Each Strategy

	Strategy		
Reviewer	*1*	*2*	*3*
1	6.01	−0.13	4.12
2	3.19	−1.40	3.74
3	1.57	2.36	3.17
4	0.72	1.47	2.25
5	0.29	2.49	3.35
6	0.04	3.03	1.25
7	0.00	4.02	1.36
8	−0.11	4.15	1.89
9	−0.53	2.08	3.03
10	−0.78	0.71	4.54
Mean	1.04	1.88	2.87

Note: Strategy 1 = relative variation; Strategy 2 = large absolute changes; Strategy 3 = intervention effects.

Although the strategies are not orthogonal, these correlations suggest that those experts who attended to the absolute size of effect were also less able to differentiate among the intervention effect patterns.

DISCUSSION

Visual inference is the predominant mode of data analysis for single-subject experimental designs (Kratochwill & Brody, 1978). The prominence of the field of applied behavior analysis, which depends heavily on single-subject methodology for its research base, attests to the usefulness of visual inference. Recent research, however, has shown low agreement between statistical and visual methods (Jones et al., 1978), low interjudge agreement (DeProspero & Cohen, 1979), and the tendency for students trained in single-subject methodology to look for the absolute size of change between phases as well as the inability to differentiate among common intervention effect patterns (Wampold & Furlong, 1981). The purpose of the present study was to determine if experts in single-subject research would be able to differentiate among intervention effect patterns and if these judgments would be affected by changes in variation in the data.

The results indicated that a majority of the experts (7 out of 10) looked for common intervention effect patterns in the time-series data presented to them; however, only a minority of the experts (2 out of 10) appeared to take variation across phases into consideration when using visual inference. The correlational analysis revealed a tendency for those experts who attended to the absolute size of the intervention effects to do more poorly at the task of classifying the effect patterns. In addition, the index values shown in Table 2 suggest that among the 10 experts, four general visual inference styles were revealed. Reviewers 1 and 2 appeared to look for intervention effects while attending to relative variation. Reviewers 3, 5, and 9 attended to the intervention effect patterns but also attended to

large changes in the data. Reviewers 4 and 10 sorted the graphs according to the intervention effects but appeared not to attend to the size of this effect, either in an absolute or relative sense. Finally, reviewers 6, 7, and 8 were unable to differentiate among the intervention effect patterns and looked only for large, dramatic changes from phase A to phase B. These findings are, perhaps, not too surprising since many advocates of visual inference define its value as being the identification of large, dramatic intervention effects (e.g., Baer, 1977). However, the results do suggest that apparently most experts do not systematically compare the size of the effect to a referent such as variation; consequently, graphs that are otherwise mathematically equivalent are seen as being different. These findings also confirm that the manner in which experts view graphs and interpret them is variable, thus providing an explanation for the lack of interjudge agreement reported in previous research.

Three limitations of this study should be taken into consideration. First, intervention effects and variation are only two, albeit important, factors that influence the process of visual inference. A number of other factors operate when an individual looks at a graph and attempts to draw conclusions from it. This appeared to be true for the experts in this study since they formed an average of six groups (range = 4-9) when classifying the 36 graphs presented to them. Inasmuch as the classification strategies against which the experts' performances were evaluated contained a maximum of three groups, it appears that they were making more detailed discriminations about the similarities and differences of the graphs than were represented in the strategies tested.

A second caution about the results of the present study involve the nature of the graphic stimuli presented to the experts. To manipulate and control the parameters of the graphs, it was necessary to generate hypothetical data. In some respects, this reduced the visual inference process to an artificial one devoid of important behavioral information available in research and clinical settings. Although this study did not perfectly model the use of visual inference, it did provide useful information regarding its reliability and dimensions that affect its use, which are prerequisites to making valid clinical judgments (Furlong & Wampold, 1981).

A third possible influence on the results is that the data sets contained only two phases, a research design that is infrequently used by applied behavior analysts. The decision to use two-phase data sets was made so that the graphs would complement the examples typically used in the single-subject methodological literature. Thus, the A-B design modeled in the graphs most likely decreased the difficulty of the visual inference task.

It should also be noted that each reviewer actually received a unique subset of the graphs generated by the computer program. This procedure allows for logical generalization of the findings beyond the specific graphs used in the study, something not true of previous research in this area (e.g., DeProspero & Cohen, 1979). Furthermore, this procedure ensured that idiosyncratic characteristics of a particular graph did not systematically bias the sorting task. That is, even if by chance the random data generation procedure produced a graph that visually appeared to have a convincing pattern, where none was intended, this did not significantly bias the outcome of the study.

For practitioners and researchers who frequently use visual inference, the results suggest that the following issues merit consideration. First, the construct of "variation" needs more careful definition in the methodological literature so that there will be clearer guidelines for its use as an interpretative dimension of visual inference. Second, since the change in the scaling transformation altered the experts' analysis, careful consideration should be given to graphing techniques—what looks convincing on one size graph paper may look much different on another. In sum, those who use visual inference

should be more cognizant of how scaling influences the appearance of graphs, how variance relates to the size of the observed intervention effect, and whether the up-and-down profile of the data represents meaningful patterns or random fluctuations.

The results of the present study offer only a glimpse into the process of visual inference as there are undoubtedly other dimensions that need to be addressed. It is quite possible that variables such as the evaluation of delayed effects, inclusion of mean or slope lines, the relationship of the last point in a phase to the first point in the subsequent phase, or the inclusion of graph keys may also affect visual inference judgments. As additional research is conducted, the goal should be to develop a clearer understanding of how researchers and clinicians look at graphs and visually analyze them. A better understanding of this process will undoubtedly improve the reliability of visual inference and provide behavior analysts with an increased capacity to determine if they have established experimental control over the variables in their studies.

REFERENCES

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

DeProspero, A., & Cohen, S. Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis,* 1979, **12,** 573-579.

Furlong, M., & Wampold, B. Visual analysis of single-subject studies by school psychologists. *Psychology in the Schools,* 1981, **18,** 80-86.

Hersen, M., & Barlow, D. *Single-case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press, 1976.

Hubert, L., & Levin, J. Evaluating object set partitions: Free sort analysis and some generalizations. *Journal of Verbal Learning and Verbal Behavior,* 1976, **15,** 459-470.

Hubert, L., & Schultz, J. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology,* 1976, **29,** 190-241.

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

Kratochwill, T., & Brody, G. Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification,* 1978, **2,** 291-307.

Parsonson, B., & Baer, D. The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change.* New York: Academic Press, 1978.

Posner, M. *Cognition: An introduction.* Glenview, Ill:. Scott Foresman and Company, 1973.

Wampold, B., & Furlong, M. The heuristics of visual inference. *Behavioral Assessment,* 1981, **3,** 79-92.

Received December 9, 1980
Final acceptance January 7, 1982