2 The Visual Analysis of Data, and Current Research into the Stimuli Controlling It

Barry S. Parsonson
University of Waikato

Donald M. Baer
University of Kansas

THE CASE FOR VISUAL ANALYSIS

Visual analysis is one of the oldest forms of data analysis. It is, in essence, simply the array of one set of information relative to one or more other sets of information, so that a viewer can draw a reasonable conclusion or make a reasonable hypothesis about any relationships or lack of them among these sets. Very often, the array takes the form called a graph, especially when at least one set of the information is quantitative; in other cases, especially when all the sets are qualitative, it takes the form of essentially a picture or diagram (cf. especially Tufte, 1990).

The point of the array is that the viewer can conclude or hypothesize about these relationships simply by visual inspection of the array: The viewer can *see*—not read, deduce, or derive, but see, and see quickly—the relationship or its absence. Thus a map shows where one place is relative to another; the relationships to be seen almost immediately are the distance and direction of any place from any other place. Certain maps show an additional relationship: where one place is relative to some means of getting there from another place. On one map, that additional element may be walking paths; on another, roads; and on another, airline terminals. The relationships that emerge from such maps are how to get from here to there by one or another means of transportation, and how long it will take—and occasionally the wry fact that you can't get there from here by that means.

The fact that visual display and analysis is one of the oldest forms of discovery and communication does not signal its obsolescence, not even in an age of ultrahigh technology. Indeed, some of our latest technology is dedicated to

maximizing dramatically the efficiency of constructing and disseminating visual analyses; these are the modern computer graphics programs. Yet in behavioral research, which is the context for this book (and this chapter), by far the prevailing mode of analysis is statistical, not visual. The multivariate analysis of variance, culminating in a table not of data numbers but of numbers testifying primarily to the probability that the data patterns could have arisen by chance, is far more frequent than a picture of the underlying data numbers themselves. Yet we can have such a picture, often by the construction of lines or other geometrical forms to show the relationship of some behaviors to the variables that may control or otherwise relate to them.

Thus, the purpose of this chapter is to state again the case for the visual analysis of behavioral relationships through graphs, and most especially for the outcome of experiments in which an ongoing, repetitive behavior is altered in its time course by the deliberate, repetitive alteration of one or more of its environmental conditions. In that context, there are at least six advantages to be gained through graphic analysis:

1. It is visual, and thereby quick to yield conclusions and hypotheses.
2. Graphs can be quick and easy to make with no more technology than grid paper, pencil, and straight edge. However, if the latest computer graphics technology is to be used, then speed and ease are recaptured only after an initial high cost of money, time, and training.
3. Graphing comprises a remarkably wide range of formats, even outside of the latest computer graphics technology.
4. Graphed messages are immediately and enduringly accessible to students at unusually diverse levels of training.
5. In representing the actual data measured, graphs can and usually do transform those data as minimally as possible. In those paradigms of knowing wherein the measurable data under study are the reality to be understood (cf. Heshusius, 1982, for a presumably different paradigm), that is an obvious virtue.
6. The theoretical premises underlying graphs are minimal and well known—that what we are interested in can be made visual, and that almost all of us are skilled in responding to visual isomorphisms of the world in ways that make the world useful. By contrast, the theoretical premises underlying the defensible use of statistical analysis are numerous, complex, diverse, and frequently arcane to the majority of their users. Thus, statistical analysis users find themselves relying on techniques subject to apparently endless debate about their suitability for given problems—a debate often accessible to only a small minority of the users. Graphs do not present us with an estimate of the probability that the patterns and distributions of the data we have gathered could have arisen by chance if the variables coupled with them are in fact not functional for them. Instead, graphs invite us to make that judgment ourselves (as well as many others exemplified herein). Thereby graphs create two audiences among researchers: (a) Some researchers will see that kind of judgment as merely a personal one. They seek a science based on objective rather than subjective judgments, so they must go further than graphic analysis in their search for an apparently objective estimate of the probability. Statistical analysis will seem to offer it to them—until they have studied its workings and underlying assumptions thoroughly enough to see how many essentially personal judgments they must make in order to estimate the suitability of each of the many models of statistical analysis for their problem and its data. (b) Some other researchers will prefer to make their own judgments about the patterns and distribution of the data; they will compare those judgments with interest but not submissively to the judgments of their peers. They will be glad to have all the data as accessible as possible, especially as simultaneously accessible as possible, and will avoid any technique that makes a decision for them in ways that are exceptionally difficult to inspect in its reduction, transformation, collation, and integration of each data point into some final decision.

To illustrate many of these points, suppose that we want to know if a certain well-measured but momentarily unnamed behavior, B, is affected differently by two well-controlled but unnamed environmental conditions, Condition 1 and Condition 2. We can experimentally alternate Condition 1 and Condition 2, each for varying lengths of time, holding everything else as stable as we can, and meanwhile measuring behavior B steadily and repeatedly under each repetition of each condition. If we do that, we can then compare the level of the behavior typical under the repeated applications of Condition 1 to the level of the behavior typical under the repeated applications of Condition 2. Table 2.1 shows every measurement of B in each of a mere two alternations of Conditions 1 and 2.

Certainly there are some clear messages to be derived from these numbers, testifying to a possibly complex differential relationship of the level of this behavior to Conditions 1 and 2. The interesting question here is only how long it takes the reader to extract and appreciate all of that message.

Now, here is a graph of the data in Table 2.1.

This graph shows us *at a glance* that Condition 2 produces more of this behavior than does Condition 1, and that initially it does so not all at once but gradually, as if a new skill were being shaped, whereas the second time that Condition 2 is applied, it produces its effect immediately, as would be expected from stimulus-control effects rather than reshaping. The possibility that the first application represents shaping and the second stimulus control is perhaps not clear enough to be affirmed flatly by every viewer, but it is clear enough to suggest further investigation of that possibility. The graph also shows us at a

TABLE 2.1
Hypothetical Data Gathered Under Two Conditions

Condition 1	*Condition 2*	*Condition 1*	*Condition 2*
1	5	4	11
1	6	2	10
2	6	0	11
0	8	2	9
1	7	1	9
2	9	1	8
1	8	2	7
0	11	0	7
1	10	0	9
1	10	1	8
	11	2	7
	9	1	6
	8		7
	8		5
	9		5
	8		
	10		

glance how much more of the behavior is produced by Condition 2 than Condition 1. It also suggests that the ability of Condition 2 to make this change in the behavior is perhaps decreasing by the end of its second application. That possibility, too, is perhaps not clear enough to be affirmed flatly by every viewer, but it, too, its clear enough to investigate further.

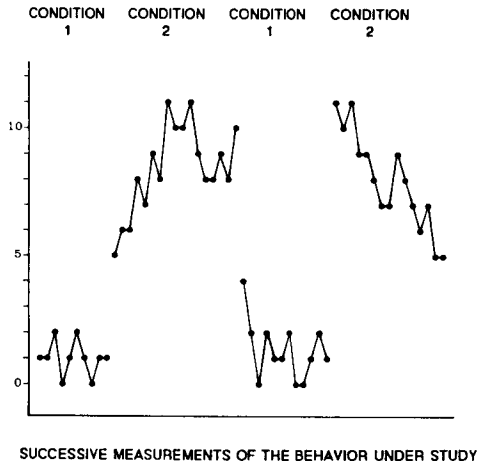

FIG. 2.1. A graphic representation of the hypothetical data of an ABAB single-subject experimental design offered in Table 2.1.

Of course, all of those conclusions and hypotheses can be extracted from the preceding table as well—the data are identical. The reader should ask how quickly and how fully both the clear and possible relationships of B to Conditions 1 and 2 were extracted from the table, relative to how quickly and how fully they were extracted from the graph. Each method of presentation offers us all the information that we collected; the table offers it to us as an array of numbers relative to their controlling variable, and the graph offers it to us as a picture of those numbers in that same array. The picture is a transformation of the numbers, certainly, but it is a remarkably minimal transformation of them. It still leaves the judgment of whether this behavior does indeed relate differentially to Conditions 1 and 2, and if so, exactly how, entirely up to us, the viewing researchers—the people who want those answers, and have been trained as best we know how to find them out.

Another interesting question is whether the viewer can decide whether the Condition 1 distributions of B are no more different from the Condition 2 distributions of B than could reasonably have arisen by chance if Conditions 1 and 2 are not functional for B. A fair variety of statistical analyses would assure these viewers that the differences in these two distributions could hardly have arisen by chance; a few other statistical methods would suggest to them that they could. Experienced visual analysts (we daresay) will typically decide fairly quickly that these differences are certainly not chance differences, and inexperienced visual analysts will come to the same conclusion even more quickly, interestingly enough. In our experience, they tend to respond to the entire picture as a clang! phenomenon, and not go on to examine the time course of each data path, point by point, and then reconsider them as a pattern of four aggregates, two with quite different internal trend patterns (the two Condition 2s), one without (the first Condition 1), and one either without or with a very short-lived one (the second Condition 1). But both kinds of visual analyst will decide that the conditions are functionally different for B far earlier than a statistical analysis could have agreed or disagreed with them. The experienced visual analysts will long since have gone on to consider those other possibilities in how B relates to Conditions 1 and 2; some of those analysts will affirm that the trends are real, and others will suspect (hypothesize) that they are but ask for more data before affirming that they are. Indeed, experienced visual analysts, studying their graphs as each point was added to the prior ones, would probably not have ended either of the Condition 2s when they were concluded in this hypothetical study: Those only hypothesizing that the trends were real would have gathered more data so as to be sure; those who were sure would have gathered more data to see how the trends would have ended (for there is always good reason to suppose that they will not continue as linear trends, especially on a graph with an arithmetical ordinate).

Graphic analysis is usually fast, clear, and easy; but it is important to consider why it is, so as to appreciate when it will not be. Graphic analysts often have the puzzling experience of finding normal adults who not only say that they cannot

read graphs, but behave in a variety of other ways that confirm their premise. Yet most of us can see effortlessly when the lines of a graph, despite their local irregularities, as a whole are at different levels, and slant upward or downward or are horizontal, or change systematically from one of those states to another. Where did we learn that? One would think that these are easy generalizations from the facts that sighted people have already learned to appreciate–at a glance!–in the rest of their world. After all, their histories with lines are vastly more extensive than their histories with numerals: They constantly scan the world within which they move for the height, rise, fall, and stability of its contours, on which they must stand, sit, walk, ride, push, pull, drive, and place or retrieve something, and over which they will climb or jump. The "horizon" of "horizontal" is a key part of their eternal stimulus controls for where they are, where they are going, and for their balance. By contrast, their behavior with numerals is a mere occasional hobby.

On the other hand, sighted people read graphs quickly and easily only to the extent that they generalize from their extensive training in a real, three-dimensional visual world to the relatively new two-dimensional visual world of graphs; and behaviorists have learned that failures of generalization are never oddities. Of course there will be people who do not immediately and easily read in graphs everything that the graphs offer. By contrast, we learned to respond to numerals in the same domains in which they are used to represent data and their statistical analyses—on pages, chalkboards, and monitor screens. Numerals present no problem in setting generalization.

Thus, it is clear that despite the advantages of graphic analysis as a quick, easy, open, and widely accessible process, it is also true that both graph construction and graph reading are skills to be acquired. In recognition of that, a chapter like this in past years would sketch some skills from both of those skill classes. However, in 1983 and 1990, Edward R. Tufte published two classic texts on graphics: *The Visual Display of Quantitative Information* (1983) and *Envisioning Information* (1990). In our opinion, those texts thoroughly and engagingly display virtually everything there is to offer so far in constructing clear and evaluative graphics for both quantitative and qualitative data, and their relationships to controlling and other variables. Consequently, we earnestly recommend the reader to them both, and turn here to a consideration of recent research into the reading of graphs.

SOME VARIABLES CONTROLLING GRAPH READING

The past decade has shown increased attention to visual analysis, both as a skill to be taught in textbooks, especially behavior-analysis textbooks (e.g., Cooper, Heron, & Heward, 1986; Kazdin, 1982; Tawney & Gast, 1984), and as a skill to be analyzed, or at least assessed. Most of that research is stimulus-analytic; it is

aimed at uncovering the stimuli that currently control visual analysts' judgm of what the data show, even if not illuminating the source or the modifiabilit those judgments.

Thus, research to date has focused on the extent to which visual analysts see the major kinds of effects that can be expected from experimental inter tions into an ongoing baseline of behavior that can be graphed as a time cou simple changes from baseline in the mean level of the behavior under stu usually referred to as *mean shift*; changes in which each successive data poir an interactive function of the intervention and the level of the immediately p data point(s), yielding at least changes in trend and sometimes simultane changes in both level and trend, and often referred to as *serial depende* degrees of such changes, simple or interactive, relative to both baseline intervention variability, thereby affecting the degree to which the ranges of baseline and intervention data points overlap; and changes only in degree variability. Research has also considered a few factors orthogonal to the beha change accomplished by the intervention, such as different graphing techniq for example, adding trend lines (calculable of course in a variety of ways) to data path(s), or choosing logarithmic rather than arithmetic scales for the o nate. Most of these topics will be discussed.

Serial Dependency

A pioneering study by Jones, Weinrott, and Vaught (1978) found that m agreement between visual and time-series analyses was both low (60%) a what little there was of it, inversely related to the degree of serial dependenc the data—that serial dependency decreased agreement between the two meth of analysis. This study also pioneered in two frequently unexamined, someti implicit assumptions: (1) That it is relevant to compare the seemingly unfor and often quite variable, judgments arrived at through visual analysis to supposedly uniform judgments imposed by statistical analysis, as if the la were somehow a standard—indeed, as if it were the truth; and (2) That unif mity across visual analysts in coming to their conclusions is a virtue.

Neither of these assumptions has much currency in modern visual analy (1) There is no single process called *statistical analysis* to provide that putat standard or truth; there are many of them, which in their diversity often all quite different conclusions. When that happens, their users must finally mak judgment about which of them is "best" for the current problem, a judgment t frequently proves unreliable across the scientists considering the problem (as most cursory examination of that literature will show); thus a supposedly obj tive, uniform process becomes instead a matter of individual scientific judgm But that is exactly the point that visual analysts make about the behavior of do science: (2) One of the central points of visual analysis is to leave with maxi clarity to each researcher and each member of the research audience any ju

ment about whether the current experimental intervention has had an effect on the behavior under study. Perhaps that is done in honor of the frequency with which science has been advanced by the disagreements of a few data analysts with the prevailing conventional wisdom about what the current data mean. One or both of these typically unexamined assumptions operates in virtually every study that will be reviewed here.

For example, to evaluate different types of serial dependency, Rojahn and Schulze (1985) used computer-generated AB-design graphs showing no, weak, or strong serial dependencies, some of the moving-average type and some of the autoregressive type; in addition, the graphs were constructed to represent five different significance levels of treatment effect (p = .50, .10, .05, .01, .001). These researchers then asked judges to rate 70 such graphs on a 5-point scale made up of just those probability values. The results did not support those of Jones et al. (1978), in that serial dependency did not much affect agreement between visual and statistical analyses. Rather, it was found that the more pronounced the moving-average and autoregressive effects, the greater the agreement between the two modes of analysis. The moving-average and autoregressive processes affected this agreement somewhat differently; strong autoregressive processes in particular led judges to overestimate treatment effects relative to statistical analysis.

Studies by Ottenbacher (1986) and Gibson and Ottenbacher (1988) investigated the effects of six graphical characteristics, including serial dependency, on interjudge agreement. Ottenbacher (1986) asked 46 occupational therapists to indicate if change had occurred between phases on five specially devised AB graphs. Gibson and Ottenbacher (1988) obtained ratings (0–5) of significance of change in 24 AB-design graphs from 20 rehabilitation therapists. Interjudge agreement was not affected by serial dependency in either study. Gibson and Ottenbacher (1988) also found no effect on the confidence in their decisions reported by the judges. However, these two studies appear to have confounded autocorrelation and mean shift, especially the Gibson and Ottenbacher data, in which the two variables intercorrelated at 0.73. Thus, judges may have confused high serial dependency with intervention effects. This would be consistent with the results of the previous studies. Still, one can always ask for further investigation into the relationship between mean shift and serial dependency. The amount and effects of serial dependency existent in applied behavior-analytic data are still much debated (Baer, 1988; Busk & Marascuilo, 1988; Huitema, 1986, 1988; Sharpley & Alavosius, 1988); thus it remains difficult to evaluate as a frequent "threat" to visual analysis.

Mean Shift

Pattern and degree of mean shift were among four data-path variables DeProspero and Cohen (1979) varied in the specially produced ABAB graphs they

submitted for judgment to 108 editorial board members and guest reviewe two behavior-analytic journals. Their graphs represented three patterns of n shift: "ideal," in which the phase means changed as "expected" with "ex mental" conditions; "inconsistent," in which the mean shifted only in the fin phase; and "irreversible," in which the mean shifted only in the first B phase remained stable thereafter. In addition, three degrees of mean shift were played, and while this varied from graph to graph, it was held constant w graphs. In this study, the pattern of mean shift proved to be critical, on average, in that mean ratings of "experimental control" were high only f least some of the "ideal" graphs showing large degrees of mean shift. A st tical analysis revealed that pattern and degree of mean shift were highly sig cant main effects, yet together accounted for only a small proportion of the variance. (It might be interesting, and certainly would be democratic, if sin analytic effort were invested to determine the extent to which highly signific levels and proportion of variance accounted for are controlling variables in conclusions of given populations of statistical-analysis consumers, especiall that typical publication greatly emphasizes the former over the latter.)

Knapp (1983) used AB graphs with a mean value of 5 in baseline and a r of nine mean values between 2 and 8 in the B phase, to see how that would a the judgments of three different groups: that much-studied subpopulation. editorial-board members of two behavior-analytic journals, as well as grad behavior-analysis students and undergraduate psychology majors. A statis analysis revealed significant main effects for mean shift, graph type, and interaction. Extreme (e.g., 5-to-8 and 5-to-2) or zero (5-to-5) mean shifts judged similarly regardless of graphing technique. At more moderate level mean shift (e.g., 5-to-3.5 or 5-to-6.5), graphing technique became critical. J ments generally were comparable for those types of graphs commonly evalu by applied behavior analysts, namely, arithmetic-ordinate and abscissa cl with a space or line between phase changes.

Ottenbacher's (1986) five graphs (discussed earlier) had also varied in de of mean shift. Most judges saw "significant change" in the three graphs with highest mean shift (which incidentally had the highest autocorrelation co cients). Interjudge agreement about change was higher when the mean sl across phases were large. Gibson and Ottenbacher (1988) (discussed earlier) also included degree of mean shift as one their six variables, and again l mean shifts yielded higher interjudge agreement and greater confidence in j ments.

A major problem in studying the detection of mean shift is confoundi with other graphic characteristics, which may well be why it accounts for so l variance in judgment while emerging as a statistically significant main effe the DeProspero and Cohen (1979) study. Mean shift will always accom between-phase changes in trend and/or level, and may accompany change variability. And sometimes it is confounded with serial dependency (Gibso

Ottenbacher, 1988; Ottenbacher, 1986). Mean shift will be present when there is no change in level or trend and no intervention effect between phases in the case of an upward/downward baseline trend continuing into an intervention attempt. In this case, a visual analyst relying only on mean shift as an indicator of an intervention effect will report change quite frivolously. Experience suggests that this misjudgment is more likely to be made if variability in the data paths masks perception of the absence of change in level and/or trend, or if phase means lines, emphasizing the shift, have been added to the graph. One way of minimizing such errors of judgment is to use trend lines (Baer & Parsonson, 1981).

Level and Trend

Half of the graphs generated by DeProspero and Cohen (1979) had a 30° upward trend, the remainder had zero slope. Their statistical analysis did not identify slope as a significant variable, although the data in their Table 1 reveal that judges' ratings almost always indicated less "experimental control" in graphs with that slope than in equivalent graphs without it. This points to some visual analysts responding to absence of trend as indicative of control. Indeed, it was the criterion most frequently mentioned by the judges.

In a study of teachers' abilities to discern trends, Munger, Snell, and Loyd (1989) investigated the effects of ascending, descending, flat, and flat but variable data paths, and four different weekly frequencies of probe-data collection (1, 2, 3, or 5 days per week). The teachers rated the degree of student progress in reading accuracy that they could see in the graphs, and also made mock decisions on program continuation. The graphs were derived from student records representing each of the data-path trends, modified to show different probe frequencies by removing enough of the intervening data points. In general, progress judgments showed accurate trend discrimination. Program-decision data showed that the average teacher would continue programs with ascending data paths and would recommend changing those with descending, flat, or flat but variable data paths. Changing the number of weekly probes did not affect judgments of progress on ascending data, but judgments were inconsistent for graphs with the other three types of data paths. Program decisions were affected by frequency of data collection (by the number of data points available for the decision) for all four classes of data path. With a very short baseline (five data points) and seven probe points spread over the 55 intervention days on the once-a-week probe graph, trend estimation is of course quite vulnerable to one or two disparate points (Cleveland, 1985). Different judgments might have been produced by plotting the dependent variable as a function simply of successive probes, rather than including in the visual display the number of days from start of baseline.

Judgments of changes in both level and/or trend were studied by Wampold and Furlong in two studies (Furlong & Wampold, 1982; Wampold & Furlong, 1981). They started with three prototypical AB graphs showing, respectively,

changes in level, trend, and both level and trend. They transformed these graph in three ways (referred to as standard, scaling, and variation transformations) t modify the visual appearance of these data paths. The judges were a group o graduate students studying single-subject research methodology, a group of stu dents studying multivariate statistics (Wampold & Furlong, 1981), and the inev itable sample of editorial-board members of a behavior-analytic journal (JABA (Furlong & Wampold, 1982). These judges were asked first to sort 36 graphs int whatever number of groups were justified as showing "similar effects" differen from the effects of the other groups; and then to identify the common feature(s of the graphs in each group. The single-subject students often responded pri marily to the absolute size of the change from A to B; they were influenced mos by those scaling transformations that enhanced both the variation and the size o the intervention effect. The multivariate students were mainly influenced b changes in level and/or trend, which often were not discriminated as such. Th JABA reviewers typically did discriminate changes in level, trend, and level plu trend, and also attended frequently to absolute size of effect. The researcher suggested that the single-subject students and JABA reviewers were so strongl influenced by size of effect, rather than by more subtle relative variations in th data, because the identification of large changes has been emphasized as crucia to the character of applied behavior analysis (e.g., by Baer, 1977). (However they drew this conclusion without evidence that their judges had ever even rea those arguments, let alone agreed with them.) Ottenbacher (1986) also found tha the detection of trend changes as such was made moderately difficult by th judges' ideas about clinically significant changes; the relevant correlation wa 0.59. This finding was replicated by Gibson and Ottenbacher (1988), who ob tained a similar correlation of 0.64, and also found raters' confidence in thei judgments was lowest on graphs with trend changes across phases; they con cluded that it was the characteristic most often associated with inconsisten interpretations.

The effects of level changes were studied by Gibson and Ottenbacher as well they correlated negatively with interrater disagreement (−0.59), rater uncertaint (−0.45), and the raters' confidence in judgment (−0.52). The authors concluded that level change is the characteristic associated with the highest degrees of rate agreement, rater certainty, and rater confidence. Similarly, Bailey (1984) ha special-education graduates of a course in applied behavior analysis judg whether or not significant between-phase changes in level or slope had occurre on arithmetic and semilogarithmic versions of five graphs previously used b Jones et al. (1978). Interjudge agreement on changes in level and trend o unmodified graphs (i.e., without trend lines) was consistently lower for tren changes.

The findings of Wampold and Furlong (1981), Gibson and Ottenbache (1988), and Bailey (1984) agree: Change in level is more often agreed on tha change in trend. The notion that slope can be difficult to judge is also supporte

by the work of Cleveland and his colleagues (Cleveland, 1985; Cleveland & McGill, 1987a; Cleveland, McGill, & McGill, 1988). In what may be a recapitulation of the Weber–Fechner Law, their subjects' judgments of slope depended heavily on the magnitude of the relative angle that the data paths create when graphed, and that how steep an angle differential is needed to be seen as such varies widely across judges. Thus, research into ways to increase accurate discrimination of relative angle differential, independent of their absolute magnitudes, is relevant. The use of trend lines for that purpose is reported later.

So far, the investigation of trend-change detection skills has focused on abrupt, sustained changes between phases. The detection of delayed or temporary changes, especially within-phase changes, which also are relevant to visual analysis of behavioral data (Parsonson & Baer, 1978, 1986), remains unanalyzed. Similarly, the study of level-change detection has investigated either change in overall phase level (Furlong & Wampold, 1982; Wampold & Furlong, 1981) or abrupt change between the last data point in one phase and the first data point in the succeeding phase (Gibson & Ottenbacher, 1988). Delayed and temporary within-phase level changes (Parsonson & Baer, 1978) are also important real-world processes, but would not be caught (even if present) by the change-of-level judgments required in the studies reviewed so far. This is of course not a criticism of those studies, only a reminder that research examining judgments of within-phase changes, and their interaction with between-phase changes, is needed for a more complete understanding of visual analysis.

Variability and Overlap

The effect of variability in the data path has seen frequent study. DeProspero and Cohen (1979) included within-phase variation as a major variable. Absent trend, lower variability made the detection of experimental control more likely, but only in the "ideal" graphs (described earlier) and even there, only if the mean shift was of a suitable kind (p. 576). Of course large, stable mean shifts and low variability in ABAB patterns yield accurate detections of the effects that the researchers have programmed into the graphs to be seen. Indeed, control of variability is one of the goals of laboratory analysis, where it is often possible (Sidman, 1960); clear attainment of that control is usually indicative of thorough, correct analysis of the usually numerous controlling variables. In application, it often is possible to attain control of only one or two of those variables, and if they are not powerful, the effects of doing so will often escape detection. Thus, studies like these, which sometimes simulate the data of situations in which only one or a few variables have been brought under control, have great relevance for application. They also suggest that those visual analysts whose standard tactic is first to achieve thorough control of the variability of their baseline, and only then to introduce an experimental intervention, may well evaluate graphic data rather differently from those researchers who have rarely if ever had that possibility.

Perhaps that is why Ottenbacher (1986) found that changes in variab across phases did not cause very much disagreement between judges al significant changes between A and B phases of the graphs they viewed. Gil and Ottenbacher (1988) replicated the finding, and again found only a w correlation (0.41) between change in variability and the average judge's eva tion of the graphs. Unfortunately, that analysis did not reveal whether th judgments were influenced by the direction of the graphed change in variabi Is a variable "baseline" becoming a stable "intervention" different from a st baseline becoming a variable intervention? In basic analysis, the two cases s be symmetrical; in applied research, they probably would not often be seer that. The rehabilitation therapists serving as judges in this study may not h seen such cases as symmetrically as the "expert behavior analysts" of the Del spero and Cohen (1979) study may have done.

The graphs of the Wampold and Furlong (1981) and Furlong and Wamp (1982) studies also included a variability transformation in which between-ph changes in level and/or trend were nonsignificant. They found that beha analysts (both single-subject methodology students and JABA editors) separa these variability transformations less often than did the graduate students stu ing multivariate analysis, and that none of the groups made consistently so judgments in the presence of enough variability (i.e., none grouped variabi transformations separately as showing no effect). Thus, it appears that w variability did not greatly affect their judgments of change in level and/or tre it still may have masked some intervention effects.

Furlong and Wampold (1982) suggest that graphs that are otherwise ma matically equivalent may be seen as different because judges fail to compare of effect with variability in any systematic way (as the analysis of varia automatically does, and as its students probably learn to do implicitly, to extent that they understand the technique that they are studying). After variability increases the number of slope and angle judgments necessary judgment, both of which are associated with lower levels of accuracy in terpretation of quantitative information (Cleveland, 1985). The data presented Munger et al. (1989) make the same point. When their teachers rated graphs v relatively few probes that showed variability but no trend, on the average t saw no student progress, and called for program changes. When probe freque was increased, however, more teachers saw some progress and considered c tinuing the mock program. If general, these findings suggest that seeing absence of trend in variable data is more difficult with many data points (i with many angle and slope judgments to make and integrate into a whole patte than with few. Once again, research into angle discrimination seems rec mended, much as Cleveland et al. (1988) have done in their investigations orientation resolution and slope judgment. Any analogous effects of adding tr lines and smoothing data variability (e.g., plotting moving averages) also n study.

Overlap between the data points of adjacent phases has not been studied much. Gibson and Ottenbacher (1988) included it as a variable, and found that it had little influence on rater disagreement, but was weakly correlated (0.36) with the raters' uncertainty (the greater the overlap between baseline and intervention data, the lower the certainty of change). Overlap also was moderately correlated (−0.74), negatively, with the detection of between-phase changes.

Of course, there are no guidelines on how much overlap is too much to allow a conclusion that the intervention has produced a change. Perhaps applied judges will see early overlap as less contradictory of an eventually useful intervention effect than enduring or later overlap—that describes much of their work eventually taken as good enough. Ultimately, the applied evaluation of any difference, including overlapping ones, depends on a cost-benefit analysis; in the case of overlap, the question changes only a little, into asking whether the benefit of changing only some data points from the baseline range is still worth the cost of doing so. The answer obviously could be either Yes or No, depending on context.

Types of Graphing

In one of the few studies of its kind, Knapp (1983) investigated how graphing techniques could affect the detection of change, using three formats: the cumulative, semilogarithmic, and frequency-polygon types. (The last term refers to the modal arithmetic-scale line graphs of everyday journal use). In addition, the study incorporated three ways of representing the AB change (on the arithmetic frequency-polygon graphs): by a space between the A and B data paths, by a vertical line between them, or without separation—a continuous data path from the start of A to the end of B. Furthermore, various degrees of mean shift between phases were represented.

The judges comprised three groups: 12 behavior analysts per group, described as undergraduates, graduate students, and postgraduate reviewers (i.e., the inevitable editorial-board members/guest reviewers). They judged a total of 147 graphs, most of them generated by the author, mostly in the AB format, and some of them taken from those used by Jones et al. (1978). The three groups of judges did not differ significantly. A statistical analysis revealed some statistically significant ($p = .05$) differences due to graphing technique, the degree of mean shift, and their interaction. Semilogarithmic charts produced the least consensus, but only on "no change" judgments; line graphs with no visual separation of the A and B data points produced the most. It was of course mainly in the middle range of mean-shift amounts that judgment differences due to graphing techniques emerged. The author concluded that the graphing technique used influenced judgments of change at critical mean shifts.

These outcomes are not surprising, given that apparent slope and apparent size of mean shift are likely to vary between arithmetic, cumulative, and semi-

logarithmic charts, with the differences becoming more critical at mode rather than extreme, mean shifts. Once again, the judges' abilities to discrimi angle differentials independently of angle magnitude may be critical. Knapp the influence of connecting the baseline and intervention data paths as an "ir vant structural feature" (p. 162), and argues that it should not affect judgn but perhaps it should not be discounted (cf. Cleveland, 1984; Parsonson & I 1978). Graphing technique may prove powerful; the Knapp (1983) study, an Bailey (1984) study to be reported in the next section, are too few to all reliable judgment.

Trend Lines

In an effort to improve the discrimination of the angles that represent s changes, especially in the face of extreme data-path variability, a numb authors have used trend lines as judgmental aids (e.g., Baer & Parsonson, I Kazdin, 1982; Parsonson & Baer, 1978; White, 1973, 1974; White & Ha 1980). Some recent studies have examined the effects of superimposing t lines generated by the "split-middle" (White, 1974) and least-squares regre procedures. Bailey (1984) obtained judgments from 13 special-education gr ate students on the significance of the change in level and/or slope in each p of five graphs (from Jones et al., 1978); these were presented both in arithn and semilogarithmic form, and with and without split-middle trend lines. trend lines increased interjudge agreement about level and trend changes in arithmetic and semilogarithmic charts. However, while judgments not simp change but of significant trend changes increased with arithmetic charts (51% to 77%), they declined with semilogarithmic (from 45% to 31%).

Clearly, the two kinds of graph can look quite different. Lutz (1949) sug that many untrained judges simply misread semilogarithmic charts. K (1983) later made a similar argument, citing the typical unfamiliarity of ju with the special characteristics of the logarithmic ordinate. Bailey (1984), commendably, did not compare his judges' accuracy in judging level or t changes against any of the numerous statistical criteria available to hin anyone); but he also did not investigate whether serial dependency affected t judgments differentially according to the kind of ordinate on which they graphed, or the use of trend lines. Those problems thus remain unstudied.

True, the Rojahn and Schulze (1985) study (already described above i serial-dependency section) did show that adding trend lines can increas similarity between visual and statistical analysis; but no attempt was ma determine how trend lines might have affected rater consistency. However, was one of the aims of a study by Skiba, Deno, Marston, and Casey (1989). asked several special-education resource teachers to judge the reading-pe mance graphs of four of their students, specifically to say whether one o interventions shown in the graphs was better than the other, to rate their c

dence in that judgment, and to state the criteria they had used (level, trend, and/or variability). Their judgments were recorded twice, once prior to training in the use of the quarter-intersect trend-line procedure (White & Haring, 1980), and again afterward. Interrater agreement increased from 0.56 at pretraining to 0.78 after learning to use trend lines. After training, the teachers showed increased confidence in their judgments, and almost total reliance on trend, ignoring level and variability as criteria. Hojem and Ottenbacher (1988) compared the judgments of a group of health-profession students after one lesson in visual analysis ($N = 20$) with those of a group given similarly brief training in computing and projecting split-middle trend lines from baseline through intervention ($N = 19$). Five of the graphs used earlier by Ottenbacher (1986) were rated for significance of change in performance across the two phases. Statistical analysis revealed significant differences in the ratings assigned to four of the five graphs by the visual analysis and trend-analysis groups: The trend-analysis group showed greater confidence in their ratings; the visual-analysis group showed slightly lower overall agreement. Commendably, Hojem and Ottenbacher did not compare their subjects' judgments to the particular statistical criteria that Ottenbacher had provided in the earlier (1986) study. If made, those comparisons would have suggested that most of the visual-analysis group judged two of five graphs "incorrectly" (not in accord with the essentially arbitrary statistical judgment). Almost all of the trend-line group misjudged one of the assumed-to-be significant graphs. On the other hand, they were not misled by the large mean shift of another graph, which the researchers' statistical analysis had suggested could most parsimoniously be considered a continuation of a baseline trend after an ineffective intervention. Most of the visual-analysis group had seen this graph as a significant intervention effect. Thus, brief training in trend-line plotting had produced somewhat more agreement with at least one line of statistical analysis than was seen in the judgments of the visual-analysis group, in both the Ottenbacher (1986) and Hojem and Ottenbacher (1988) studies.

The results of these studies suggest that trend lines through the various phases of a study (Bailey, 1984; Rojahn & Schulze, 1985; Skiba et al., 1989) or projected from baseline through intervention (Hojem & Ottenbacher, 1988) may alter interjudge agreement by providing common stimuli for determining the direction and amount of change in overall trend and/or level occurring between phases. There also is evidence from the studies by Bailey (1984), Hojem and Ottenbacher (1988), and Skiba et al. (1989) that the presence of trend lines alters judges' confidence in their decisions. In addition, it seems that adding trend lines may increase the similarity between visual and at least some statistical analyses, but not at the price of the conservatism of visual judgments. The unanswered question is whether that is a good outcome.

These findings, coupled with the difficulties of accurate slope estimation by untrained viewers, might be seen as a recommendation that trend lines always be

used. However, remember that Skiba et al. (1989) found that judges taught to use trend lines came to rely on them: They then attended much less to all other data-path characteristics, such as level and variability. Furthermore, while between-phase trend lines certainly do summarize trend and level changes instantly and clearly to the eye, they can also obscure from that same eye the within-phase changes in the variability, level, overlap, pattern, and latency of change, all of which can contribute important hypotheses about the nature of the behavior change under study (Parsonson & Baer, 1978, 1986). We need further research to show how trend-line analysis can be taught and used without paying any of that price. That research should be done under the assumption that this is only a training problem, not a fixed characteristic of visual analysis.

These studies used either median split (split-middle, quarter intersect) or least-squares regression procedures. Shinn, Good, and Stein (1989) studied the comparative predictive validity of these procedures. They asked special-education teachers to graph the reading progress of 20 mildly handicapped students. Each student's graph was offered in three versions, the first covering data points 1 to 10, the second, 1 to 20, and the third, 1 to 30. Times at 2, 4, and 6 weeks following the final data point on each partial graph were defined as the occasions when predictions based on trend-line projections from the partial graphs would be tested against actual student performance. Graduate students, well trained in the "split-middle" technique but unaware of the aims of the study, were given the partial graphs and asked to produce a split-middle trend line for each one, projecting it to the three designated prediction days. Reliability checks showed 0.91–0.99 agreement among them in generating these lines.

Then least-squares trend lines were obtained for the same data sets and projected over the same time spans. After that, actual student reading performance at each trend-line prediction point was taken as the median of the three actual data points nearest that day. Neither procedure systematically over- or under-predicted performance, but the least-squares procedure yielded better precision most consistently across all numbers of points and all three lengths of prediction.

Yet both procedures have disadvantages. In using the split-middle technique, Shinn et al. (1989) occasionally obtained very inaccurate predictions. This has also been the experience of the present authors, and the conditions that produce these deviant trend lines require examination. One contributing factor may be the requirement that the trend line must have the same number of data points on or above it as fall on or below it. Two disadvantages claimed for the least-squares procedure are its difficulty of computation, and the effects of extreme outlying data points on the trend line (Baer & Parsonson, 1981; Shinn et al., 1989). The former can be overcome by using the calculation algorithm provided by Baer and Parsonson (1981). However, the disproportionate effect of extreme outliers is inherent in the least-squares technique. Thus there is value in examining the applicability of alternative methods of generating trend lines, such as Cleveland's

LOWESS (1985) and the variety of ways of weighting regression calculations (Huber, 1973, cited in Cleveland & McGill, 1987a). The interesting question is what criteria we should use to evaluate these alternatives.

Problems in the Current Research

In all of the studies discussed here, raters evaluated only graphs. Thus, they were not evaluating data under the normal conditions of research and application.

First, there was an absence of the abundant and complex contextual information normally associated with evaluating data: the study's aims, the special characteristics of its subjects and its intervention personnel, all that is known about these intervention procedures and their interaction with these kinds of subjects and intervention personnel, all that is known about these kinds of measurement techniques and how they interact with these kinds of subjects and intervention personnel, and certainly not least, the graphing method, which in these studies may often have been different from the techniques the judges would have chosen. Indeed, the judges' additional opinions, when solicited, often enough included complaints about this (DeProspero & Cohen, 1979; Knapp, 1983).

Indeed, except for the graphs used by Jones et al. (1978), there was an absence or paucity of the usual information even on the graphs' axes—information about the dependent and independent variables, such as those considerations listed in the preceding paragraph. This point is important, in that it reminds us that there are two conceivable domains of reading graphs: the real-world domain, in which researcher's and articles' graphs always come with full contextual information and well-marked axes; and a theoretical domain in which graphs are merely abstract, content-free forms specifying that an undescribed behavior was measured by an undescribed process under some undescribed conditions over an unspecified length of time, and yielded the graph shown, much like the first graph presented in this chapter. The latter approach assumes that certain visual forms are stimulus controls for corresponding conclusions about the relationships between data and the conditions under which they were gathered—visual forms that *should* function as stimulus controls for certain conclusions no matter what the contextual information might be.

On reflection, it may seem that this approach is an attempt to convert graphic analysis, recommended here in part because it allows the researcher to exercise skilled personal judgment in deciding what an experiment shows, into the objective, supposedly judgment-free decision making put forward as the ideal outcome of statistical analysis. True, there are visual forms that tend to compel certain conclusions no matter what their content might be; the first graph of this chapter was constructed with one of those forms to accomplish exactly that. Thus, there are indeed such forms; in an earlier era, they would be referred to as *gestalts*. But the problem in relying exclusively on them is that doing so does not

represent much of the real-world research and clinical practice under study. Researchers, clinicians, teachers, and literature readers never examine content-free graphs for very long; invariably, they see those visual forms only in context, and almost certainly, their final conclusions about those graphs result from an interaction between the form, the clarity with which the actual data fit the form, and the total context relevant to those data. Thus studies of the interpretation of content-free graphs are a crucial part of the analysis of graphical interpretation, but gravely incomplete as an analysis of the real-world process.

Second, many of these studies were group designs using statistical analysis to understand the use of graphic analysis with single-subject designs. In other words, tools and designs from one domain of inquiry were used to evaluate the tools and designs of another—tools and designs that are usually adopted systematically by those researchers who have found the alternatives inadequate for their purposes. This evaluation is formally possible, of course; we remark only on the irony of it. It would be interesting to discover how many journals devoted to the statistical analysis of data from group designs would publish a single-subject design using visual analysis to clarify *their* methodology.

Perhaps a chapter like this one should confront its readers directly with a rudimentary assessment of their own visual analysis skills, now that the readers have read about the variables controlling others' visual analysis skills. For that purpose, two sets of six graphs each have been prepared and are presented here (see Figs. 2.2 and 2.3).

The first graph of the first set was constructed by entering a random-number table at random and taking the first 40 digits (the integers between 0 and 9) that the table presented there. The first graph of the second set was constructed in the same way, subsequent to a second randomly chosen entry of the random-number table. Two sets were made only because even one replication conveys a great deal more generality than a single experience.

Each of these two groups of 40 digits was then divided into the first 10 digits, the second 10, the third 10, and the fourth 10, in the order in which they were found in the table. The first and third groups of 10 were considered as an A condition and its repetition; and the second and fourth were considered as a B condition and its repetition.

The first graph in each of these two six-graph sets is simply a graph of these 40 numbers as an ABAB design. The second graph adds a constant 1 to each number in the two B conditions; the third adds a constant 2 to each number in the two B conditions, the fourth a constant 3, the fifth a constant 4, and the sixth a constant 5. Beside each graph is printed the probability-level result of what is often called an independent-*t* test of the hypothesis that the 20 As are a sample from the same population as the 20 Bs. (In that these digits in fact are drawn from a random-number table, they should be totally independent of each other—free of autocorrelation—and thus quite suitable for that test.) No other context of any sort is specified; take them as pure forms.

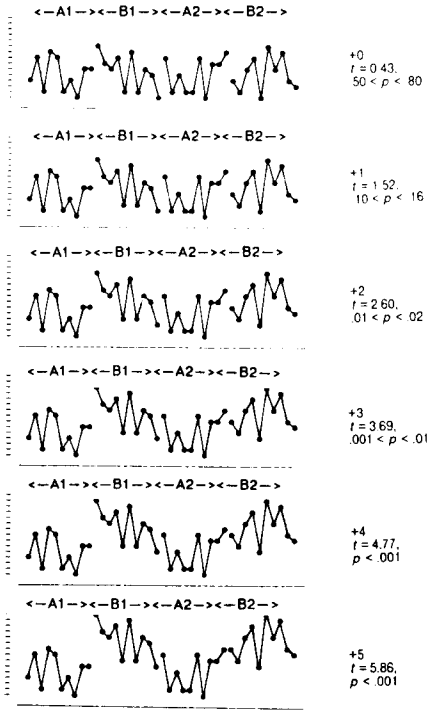

FIG. 2.2. At the top, a graph of 40 successive digits drawn from a random-number table, arranged as if in an ABAB single-subject experimental design, and below it, five regraphings of those points. In the five regraphings (the second through sixth of these graphs), a constant has been added to each of the 20 points in the two B conditions; the value of the constant is shown at the right of each graph, along with the *t* statistic and probability level resulting from an independent-*t* test applied to these data.

These graphs create three very interesting questions for visual analysts. The first question is what magnitude of a simple additive effect you require to answer Yes to either of two questions about each of the six graphs in each set: (1) Speaking only to yourself, do you think that there is a systematic B effect? (2) Are you willing to publish these data under your name as an example of a systematic B effect? Many visual analysts may discover that there is a U-shaped interaction between the size of the additive effect and the similarity of their answers to these two questions (i.e., they answer similarly with No to both questions at the +0 and +1 levels, and similarly with Yes to both questions at the +5 level, but dissimilarly to the two questions at the intermediate +2, +3, and perhaps +4 levels). Some visual analysts may also be surprised to discover that given the variability of this simple distribution, the addition even of 5 to each point in B—which on the average should more than double the mean of the B conditions relative to the 4.5 mean generally characterizing A conditions of this distribution—does not always allow them an unequivocal Yes answer to the second question.

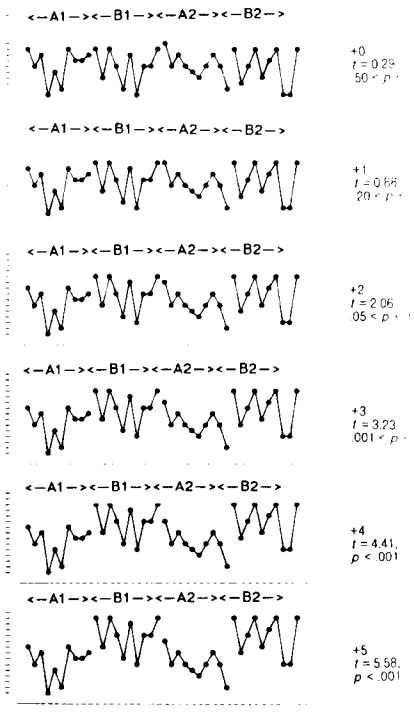

FIG. 2.3. A replication of Fig. 2.2, but with 40 successive digits drawn from another randomly selected part of the same random-number table.

The second question is whether you are as sensitive as the *t* test. Many visu analysts may find that while the addition of a constant 2 to the B points achiev (or virtually achieves) the conventional level of statistical significance via the test, they are not willing to answer Yes to even the first question until the + and some only at the +4 level or +5 level. Baer (1977) has argued that visu analysts responding to data in single-subject designs probably display low probabilities of Type-1 errors, and thereby correspondingly higher probabiliti of Type-2 errors, than would many of the usual statistical analyses, and th doing so is an important component in developing an effective technology behavior change. This tiny exercise is an example of the argument, but of cour is not a proof of its generality.

The third question also cannot be answered by so tiny an exercise, but it c be suggested by it: If you are discovering that you cannot see the effects of son actually effective interventions as reliable—in particular, the adding of 1, 2, or 4 to each observation in the B conditions—is that kind of discovery likely change what you can see as an effect in the future? In other words, could y

train yourself to detect such effects by exposing yourself to many, many graphs like these, one after another, some with an intervention of constant effectiveness, some without; then making your estimate of whether the Bs are different from the As; and then promptly being informed of exactly the extent to which the Bs are different from the As, if they are? Program your computer to present you with many such graphs of every kind of relevant effect (additive, multiplicative, autocorrelated, constant, variable, increasing, decreasing) in unknown and unpredictable order; to record your keyboarded answers to the two questions; and then to inform you of the truth. See if your standards are modifiable by this kind of feedback, which you will never encounter with the same certainty in real-life research and practice, yet which is exactly the appropriate reinforcer: knowledge of whether you were correct or incorrect. That is why graphs are made and read, is it not? Not to be rich, not to be liked, but to find out?

Third, the graphs of these studies often were devised by the researchers to display data characteristics prescribed by theories about the kinds of data that there should be, rather than representing actual data; they may have looked other worldly to many experienced viewers.

Fourth, in many instances (e.g., Rojahn & Schulze, 1985) only AB data formats were offered; these are not at all the typical designs of applied behavior-analytic research. True, the AB design is in principle the irreducible unit of single-subject research design, and some researchers under some conditions will consider it sufficient to allow a firm conclusion, when the data offer a perfect picture of experimental control. But much more of the field uses the AB design perhaps as a unit, but one so rarely offered alone as a form of proof that we could reasonably argue that for them, an ABA, BAB, or ABAB pattern is the functional unit controlling conclusions.

Fifth, at times the questions asked of the judges were ill defined, so that unmeasured, uncontrolled differences in interpretation may have accounted for low agreement. Many, many times the authors have found that students are willing to draw one conclusion in private, a somewhat different one if they expect their dissertation adviser to agree to it, yet a different one if they wish their dissertation oral-examination committee to agree, and a still different one for submission to a journal editor and, by presumption, their total professional audience. This point is not offered as an amusing picture of immature scientists at work, but as a realistic picture of virtually all scientists at work, who know not only what they think but also what other enthusiasts who like that conclusion will think, what scientists supportive of them but skeptical of their conclusion will think, and what unsupportive skeptics will think.

For example, Jones et al. (1978) asked if there was a "meaningful [reliable] change in level" across phases, but "reliable" was not defined: stable, durable, replicated? Nor was "change in level" defined (Is it the last data point in one phase compared with the first in the next, or mean shift across phases, or trend lines showing an overall change in phase level)?

Sixth, most studies included subjects with little or no knowledge or expe ence of visual analysis, and asked them to interpret the statistical or clinic significance of changes.

Seventh, in none of the studies were fine-grained analyses of within-pha variables investigated.

Eighth, in many studies a number of variables possibly relevant to visu analysis were confounded or manipulated simultaneously, making it difficult identify the specific effects of any one variable on visual analysis.

Ninth, some studies used AB graphs in which a definite effect had be systematically programmed into the B conditions, and others in which no effe other than random variation distinguished the B condition from the A conditio In these studies, it was at least possible to ask how often visual analysis cou detect the truth, which was that difference or lack of it, and to compare that ra of detecting that truth under various experimental conditions and to other met ods of evaluating the same data, such as any of the numerous statistical mod available and conceivably appropriate. But in many other studies, there was known truth; the graphs had been chosen from the journals of the field, or h been constructed to display common or tricky patterns. In this latter case, wher certain statistical analysis says that there is a difference, and another statistic analysis says that there is not, and visual analysis says that there is or is no which is correct? Indeed, does "correct" have any useful meaning in that co text? The fact that these methods often generate different conclusions about t same data is an almost useless fact, unless we know which conclusion is som how the better one.

As a consequence of these inadequacies, only tentative conclusions about t nature and effects of the variables influencing visual analysis seem justified far. The differences from actual practice mean that it is impossible to kno precisely how they affect the day-to-day judgments of persons making decisio from ongoing research or intervention data or from published research.

Future Developments

In addition to the areas of further research that have been suggested, there is need for those interested in visual analysis to become familiar with developmer in theory and research outside behavior analysis and to examine their a plicability to the data-analysis problems we face. For example, in the areas data presentation and information processing, Bertin (1981, 1983) and, as w recommended earlier, Tufte (1983, 1990) provide stimulating theoretical ide and practical methods of evaluating and presenting data graphically. Some of t most significant contributions in the area of graphic discriminations have com through the careful parametric studies of Cleveland and his associates (Clev land, 1984, 1985; Cleveland & McGill, 1984, 1986, 1987a; Cleveland et a 1988). This work provides theoretical models founded in psychophysics an

statistics that have been examined and developed thoroughly through research; consequently, they reflect a high standard of scientific inquiry and offer a conventionally sound platform for the further study of variables affecting visual analysis. We also may find ourselves shaped in our visual analyses by developments in the field of computer graphics, especially in the use of dynamic graphics (Cleveland & McGill, 1987b), to explore the effects of changing graphic formats and data-path characteristics, some of which may optimize data presentation and analysis. Finally, we can usefully explore the application of new designs and different types of graphing in our efforts to enhance communication and comprehension of the results of behavior-analytic research.

REFERENCES

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

Baer, D. M. (1988). An autocorrelated commentary on the need for a different debate. *Behavioral Assessment, 10,* 295–298.

Baer, D. M., & Parsonson, B. S. (1981). Applied changes from the steady state: Still a problem in the visual analysis of data. In C. M. Bradshaw, E. Szabadi, & C. F. Lowe (Eds.), *Quantification of steady-state operant behaviour* (pp. 273–285). Amsterdam: Elsevier/North Holland Biomedical Press.

Bailey, D. B., Jr. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. *Journal of Applied Behavior Analysis, 17,* 359–365.

Bertin, J. (1981). *Graphics and graphic information processing* (W. J. Berg & P. Scott, Trans.). New York: de Gruyter. (Original work published 1977)

Bertin, J. (1983). *Semiology of graphics.* (W. J. Berg, Trans.). Madison: University of Wisconsin Press. (Original work published 1973)

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10,* 229–242.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *American Statistician, 38*(4), 270–280.

Cleveland, W. S. (1985). *Elements of graphing data.* Monterey, CA: Wadsworth.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531–554.

Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies, 25,* 491–500.

Cleveland, W. S., & McGill, R. (1987a). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society, 150*(3), 192–229.

Cleveland, W. S., & McGill, R. (1987b). *Dynamic graphics for statistics.* Monterey, CA: Wadsworth.

Cleveland, W. S., McGill, M. E., & McGill, R. (1988). The shape parameter of a two-variable graph. *Journal of the American Statistical Association, 83*(402), 289–300.

Cooper, J. O., Heron, T. E., & Heward, W. L. (1986). *Applied behavior analysis.* Columbus, OH: Merrill.

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

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensi in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15,* 415–421.

Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of sing subject data: An empirical analysis. *Journal of Applied Behavioral Science, 24*(3), 298–3

Heshusius, L. (1982). At the heart of the advocacy dilemma: A mechanistic world view. *Exc tional Children, 49*(1), 6–13.

Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection ver trend-line analysis of single-subject data. *Journal of the American Physical Therapy Associati 68,* 983–988.

Huitema, B. E. (1986). Autocorrelation in behavioral research: Wherefore art thou? In A. Poling R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (187–208). New York: Plenum Press.

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10,* 2 294.

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

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* N York: Oxford University Press.

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

Lutz, R. R. (1949). *Graphic presentation simplified.* New York: Funk & Wagnalls.

Munger, G. F., Snell, M. E., & Loyd, B. H. (1989). A study of the effects of frequency of pr data collection and graph characteristics on teachers' visual analysis. *Research in Developmer Disabilities, 10,* 109–127.

Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from sing subject designs. *American Journal of Occupational Therapy, 40,* 464–469.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 101–165). N York: Academic Press.

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

Rojahn, J., & Schulze, H-H. (1985). The linear regression line as a judgmental aid in the vis analysis of serially dependent A–B time-series data. *Journal of Psychopathology and Behavi Assessment, 7,* 191–206.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternat perspective. *Behavioral Assessment, 10,* 243–251.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student achievement comparison of methods. *School Psychology Review, 18,* 356–370.

Sidman, M. (1960). *Tactics of scientific research.* New York: Basic Books.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subj familiarity on practitioners' judgments of intervention effectiveness. *Journal of Special Edu tion, 22,* 433–446.

Tawney, J. W., & Gast, D. L. (1984). *Single-subject research in special education.* Columbus, (Merrill.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Pr

Tufte, E. R. (1990). *Envisioning information.* Cheshire, CT: Graphics Press.

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. *Behavioral Ass ment, 3,* 71–92.

White, O. R. (1973). *A manual for the calculation of the median slope: A technique of progr*

estimation and prediction in the single case. (Working paper No. 16). Eugene: University of Oregon, Regional Resource Center for Handicapped Children.

White, O. R. (1974). *The "split middle": A "quickie" method of trend estimation* (3rd revision). Unpublished manuscript, University of Washington, Experimental Education Unit, Child Development and Mental Retardation Center, Seattle.

White, O. R., & Haring, N. G. (1980) *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.