

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

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

Furlong and Wampold (1982) suggest that graphs that are otherwise mathematically equivalent may be seen as different because judges fail to compare size of effect with variability in any systematic way (as the analysis of variance automatically does, and as its students probably learn to do implicitly, to the extent that

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

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

Trend Lines

In an effort to improve the discrimination of the angles that represent slope changes, especially in the face of extreme data-path variability, a number of authors have used trend lines as judgmental aids (e.g., Baer & Parsonson, 1981; Kazdin, 1982; Parsonson & Baer, 1978; White, 1973, 1974; White & Haring, 1980). Some recent studies have examined the effects of superimposing trend lines generated by the "split-middle" (White, 1974) and least-squares regression procedures. Bailey (1984) obtained judgments from 13 special-education graduate students on the significance of the change in level and/or slope in each phase of five graphs (from Jones et al., 1978); these were presented both in arithmetic and semilogarithmic form, and with and without split-middle trend lines. The trend lines increased interjudge agreement about level and trend changes in both arithmetic and semilogarithmic charts. However, while judgments not simply of change but of significant trend changes increased with arithmetic charts (from 51% to 77%), they declined with semilogarithmic (from 45% to 31%).

Clearly, the two kinds of graph can look quite different. Lutz (1949) suggests

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 should not simultaneously be considered a continuation of a baseline trend after

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.

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

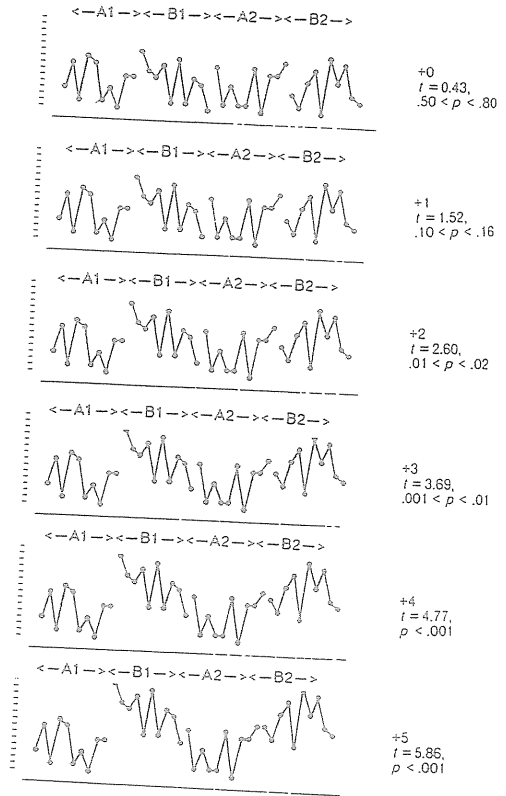
represent much of the real-world research and clinical practice under study. Researchers, clinicians, teachers, and literature readers never examine content-free graphs for very long; invariably, they see those visual forms only in context, and almost certainly, their final conclusions about those graphs 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

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.



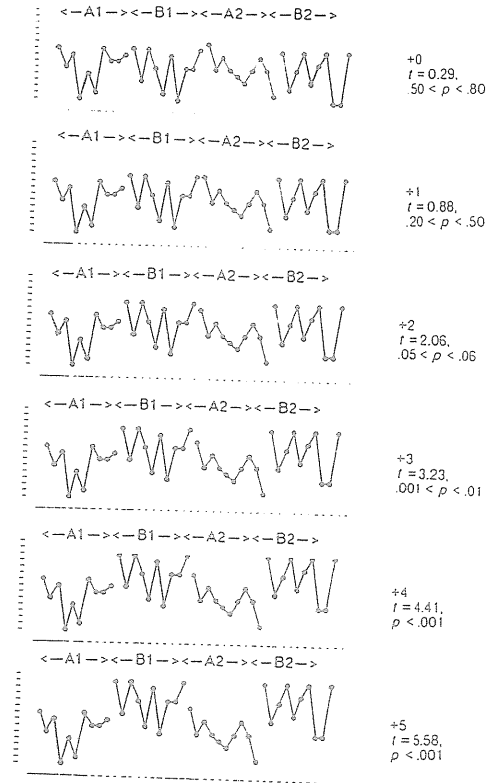


FIG. 2.3. A replication of Fig. 2.2, but with 40 successive digits drawn from another randomly selected part of the same random-number table.

The second question is whether you are as good as a random number table.

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 func-

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

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

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

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

As a consequence of these inadequacies, only tentative conclusions about the nature and effects of the variables influencing visual analysis seem justified so far. The differences from actual practice mean that it is impossible to know precisely how the effect of these variables is manifest in the actual practice of visual analysis.

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

- Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions 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 single-subject data: An empirical analysis. *Journal of Applied Behavioral Science*, *24*(3), 298-314.
- Heshusius, L. (1982). At the heart of the advocacy dilemma: A mechanistic world view. *Exceptional Children*, *49*(1), 6-13.
- Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. *Journal of the American Physical Therapy Association*, *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* (pp. 187-208). New York: Plenum Press.
- Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment*, *10*, 253-294.
- Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, *11*, 277-283.
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *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 probe data collection and graph characteristics on teachers' visual analysis. *Research in Developmental Disabilities*, *10*, 109-127.
- Ottenbacher, K. J. (1986). Reliability and accuracy of visually analyzing graphed data from single-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. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 101-165). New York: Academic Press.
- Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New

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.