analysis. That would not address the problem of reversal designs where performances are slow to return to initial baseline levels, however, unless some objective criteria were added to the strategy to allow such designs to always use the dynamic approach (see Figure 3a).

Probably the solution lies in continued examination of alternatives. Our experiences with such approaches simply do not allow us to draw firm conclusions concerning the implications of choosing one strategy over another. Before we can begin to understand what such metrics might mean we need to make direct comparisons among several alternatives and present the *real* data upon which they are based. Possibly then some accommodation between the "blind statistic" and the "experienced eye" can be found, or at least the potential dangers made explicit.

I commend Scruggs, Mastropieri, and Casto for their efforts, and encourage them and others to continue the search for meaningful single-subject outcome metrics. Such metrics would undoubtedly be of value. For now, however, I must disagree with their conclusion that the PND (or anything else) is "easy to interpret" and "reveals a consistent, meaningful outcome" (p. 31).

Owen R. White *is an associate professor in special education at the University of Washington. Working primarily with the severely handicapped, he has focused his research on methodology and the development of data-based guidelines for helping teachers to monitor pupil performance and decide if, when, and how instruction should be modified to meet individual needs.*

References

Edgington, E.S. (1980a). *Randomization tests.* New York: Marcel Dekker.

Edgington, E.S. (1980b). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2,* 19-28.

Edgington, E.S. (1982). Nonparametric tests for single-subject multiple schedule experiments. *Behavioral Assessment, 4,* 83-91.

Glass, G.V., Willson, V.L., & Gottman, J.M. (1975). *Design and analysis of time-series experiments.* Boulder: Colorado Associated University Press.

Guevremont, D.C., Osnes, P.G., & Stokes, T.F. (1986). Programming maintenance after correspondence training with children. *Journal of Applied Behavior Analysis, 19,* 215-219.

Huitema, B.E. (1985). Auto-correlation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 109-120.

Kazdin, A.E. (1976). Statistical analyses for single-case experimental designs. In M. Hersen & D.H. Barlow, (Eds.), *Single-case experimental designs: Strategies for studying behavior change* (pp. 265-316). New York: Pergamon.

Koenig, C.H. (1972). *Charting the future course of behavior.* Unpublished doctoral dissertation, University of Kansas, Lawrence.

Neef, N.A., Parrish, J.M., Egel, A.L., & Sloan, M.E. (1986). Training respite care providers for families with handicapped children: Experimental analysis and validation of an instructional package. *Journal of Experimental Analysis of Behavior, 19,* 105-124.

Schepis, M.M., Reid, H., Fitzgerald, J.R., Faw, G.D., Van Den Pol, R.A., & Welty, P.A. (1982). A program for increasing manual signing by autistic and profoundly retarded youth within the daily environment. *Journal of Applied Behavior Analysis, 15,* 363-379.

Shewart, W.A. (1931). *The economic control of the quality of manufactured product.* New York: Macmillan.

Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

White, O.R. (1984) Selected issues in program evaluation: Arguments for the individual. In B.A. Keogh (Ed.), *Advances in special education, Volume 4: Program in evaluation* (pp. 63-122).

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

White, O.R. (1972). *The prediction of human performances in the single case: An examination of four techniques.* Working Paper No. 15. University of Oregon Regional Resource Center for Handicapped Children, Eugene.

Continued from p. 42

References

Edgington, E.S. (1980). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2,* 19-28.

Edgington, E.S. (1982). Nonparametric tests for single-subject multiple schedule experiments. *Behavioral Assessment, 4,* 83-91.

Huitema, B.E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 109-120.

McCain, L.J., & McCleary, R. (1979). The statistical analysis of the simple interrupted time series quasi-experiment. In J.D. Cook & D.T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 233-293). Boston: Houghton-Mifflin.

Scruggs, T.E., Mastropieri, M.A., Cook, S.B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders, 11,* 260-271.

RASE

Rebuttal

Reply to Owen White

Thomas E. Scruggs, Margo A. Mastropieri, and Glendon Casto

The response of Owen White generally supports the notion of computing objective study outcomes for the purpose of reviewing single-subject research. He suggests, however, that our model is not sensitive to trends and proposes a model for computing percentage of nonoverlapping data based upon consideration of baseline trends. We respond that although such a procedure has theoretical merit, it has little practical utility due to the limited number of observations typically found in single-subject baselines; furthermore, when reliable and substantial baseline trends are found, interpretability of study outcomes is seriously limited.

DR. WHITE HAS delivered what we regard as a thoughtful, constructive evaluation of our efforts in quantitative synthesis. We are particularly pleased with his positive approach and his specification of possible positive alternatives to our own method. It is through such positive exchanges of ideas and information that the field is able to develop.

Dr. White (this issue) begins his paper with a note of accordance with our position that any review of literature, including single-subject literature, should be as thorough, objective, and as systematic as the studies that are contained in the review. We also are pleased to acknowledge his agreement that a standard, objective outcome metric could greatly facilitate the review process. Finally, he describes several understandable concerns regarding our percentage of nonoverlapping data points (PND) statistic, and suggests in each case positive alternatives for meeting these concerns.

First, White provides an example of a recent study in which 1 of 85 baseline data points, apparently an "outlier," appears to compromise the validity of a PND statistic. He suggests use of a computed confidence interval, or elimination of the highest 5% of baseline data points as alternatives. Second, he notes our reported lack of discriminability between outcomes that contain no overlapping data, and suggests as alternatives computation of parametric statistics or use of standardized "effect sizes" such as those used in meta-analysis of nomothetic ("group") research. Finally, he expresses concern that we have "avoided" the issue of trends and suggests several techniques for computation of baseline trends and employing them in what he terms a "dynamic" PND.

We would like to respond to these concerns in turn. Although all of White's alternatives are thoughtful and positive, none appears to be a viable alternative to our own metric. In fact, we had previously considered and rejected all of these alternatives before we began to employ the PND metric. The major problem with these alternatives is not in Dr. White's reasoning as much as the limitations of "real-world" applications of single-case investigations. We will consider these concerns separately below.

Baseline Variability

Dr. White expresses concern that the PND estimate may be compromised by an "outlier" in the baseline phase, and cites a possible example from a case with 85 baseline data points. Frankly, we wish we had encountered such problems! The reality we had to deal with was that baseline data were generally far shorter, not longer, than the "ideal." For a description of the reality of baseline data, we again recommend to the reader a study by Huitema (1985), also cited by Dr. White. Huitema calculated the actual number of baseline data presented in each of 881 baselines published in the first 10 volumes of the *Journal of Applied Behavior Analysis* (1968-1977). The median number of baseline data points reported during this period was 5; the mode was 3-4.

0741-9325/87/0082-0040$2.00

And, in fact, not one baseline reported during this 10-year period approached the baseline length described by White. Our own analyses are in complete agreement with Huitema. We are not asserting that such cases do not exist, but we hope the reader will find it understandable that we did not prespecify a convention for handling such an unusual case. If one had arisen, we might have employed a convention similar to Dr. White's "upper 5%" suggestion (we do not feel the computation of "confidence intervals" would be appropriate). However, such a convention would necessitate a baseline of 20 or more data points—a condition that occurred very infrequently in Huitema's sample and was virtually nonexistent in ours.

Finally, the issue of whether researchers are "punished" for providing additional baseline data is, as White acknowledges, an empirical question, and one that we tested empirically. In one of many confirmatory and cross-validation analyses that we did not report in our original manuscripts due to space limitations, we computed a nonsignificant (in fact, near zero) correlation between number of baseline data points and PND score. Although appropriate controls must be employed in each individual case (we certainly do not recommend a "blind," thoughtless computation of PND scores), in our samples baseline length did not compromise outcomes.

Power of PND

White points out our admission that the PND score does not discriminate between cases that produce zero overlap. One point we wish to make clear is that we did not intend our PND metric to be considered to be a measure of the *size* of the effect (as is an "effect size") as much as it is a measure of the *tangibility* (or "convincingness") of the effect. This distinction is subtle, but one which we think is important. The lack of discriminability of extreme cases is often found in nonparametric tests, and although it may lower power somewhat, it in no way compromises the overall value of these tests. Furthermore, the PND statistic, like nonparametric tests, is useful when other alternatives are not viable. In addition, it would not be possible to employ the two alternatives proposed by White without disastrous consequences. The computation of a standardized, mean-difference effect size is not appropriate for a variety of reasons, not the least of which is the reliable estimate of means and standard deviations of baselines that typically include only four or five observations (other problems are described by Scruggs, Mastropieri, Cook, & Escobar, 1986). Our experience has convinced us that such calculations result in effect sizes of enormous variability that are essentially devoid of meaning and that often directly contradict visual inspection methods, White's example notwithstanding. Finally, the suggested use of relative t and F statistics is an inappropriate interpretation of Edgington's (1980) suggestions. Edgington suggested computation of all possible permutations of t statistics *on the same data* for use in a randomization test, a possibility we discussed and discounted in our paper. Such a test requires random assignment to treatment times, a condition almost never met in single-case experiments, and results in a probability estimate only for that experiment, not on overall t or F. Edginton never recommended the use of parametric statistics *across* studies for comparison of single-case effects (he did not mention parametric statistics *at all* in his 1982 paper). Certainly the notion that such statistics are valid, although corresponding probability estimates are *not,* strains credibility.

Independence

White's final concern is that our procedures are not sensitive to baseline trends, and he offers some examples to reinforce this point. We agree that the issue of trends is an important one, and one that consumed much of our early exploration efforts. Actually, if trends or seasonal effects were to be evaluated, we would recommend use of the ARIMA model, described by McCain and McCleary (1979), rather than the simple linear regression technique proposed by White. The ARIMA (autoregressive, integrated moving average) approach allows, through computation of autocorrelations, a quantitative evaluation of serial dependence, seasonal effects, and overall trends and level changes that at first may not be apparent to the unaided eye. The problem with adapting such a model to single-case behavioral investigations is that the number of data points commonly reported is clearly insufficient for such an evaluation. If typical single-case investigations contained 20, or preferably 50, data points in baseline and treatment phases (and individual data points could be reliably estimated from graphic presentations), we would have chosen such an analysis. As White admits, however, "making meaningful predictions on the basis of five or six data points is always tenuous" (p. 38). Unfortunately, five or six baseline data points appear to be the rule, rather than the exception, in single-case investigations. In such cases, evaluation of "trends," according to White's proposed method, can produce disastrous consequences similar to those resulting from attempts to estimate means and standard deviations of such data. One such consequence is shown in the example in Figure 1, in which apparently idiosyncratic and sparse baseline data result in a meaningless trend estimate that compromises visual analysis and our own PND metric.

Another example may be found in White's discussion of possible interpretations of baseline data in his Figure 4. In describing visual analysis procedures, he refers to individuals "believing" or "considering" the possibility of meaningful trends. If such differing *opinions* are apparently possible, there are problems indeed with the evaluative criteria. Huitema (1985), in fact, did not evaluate autoregressive components (nonindependence) in baselines with fewer than six observations, and treated such data as basically unreliable.

We hold to our original argument that when "obvious" increasing baseline trends are noted, such data are basically uninterpretable, and their exclusion is analogous to the exclusion of uninterpretable studies from "group" meta-analyses. White is correct, however, that specific, quantita-

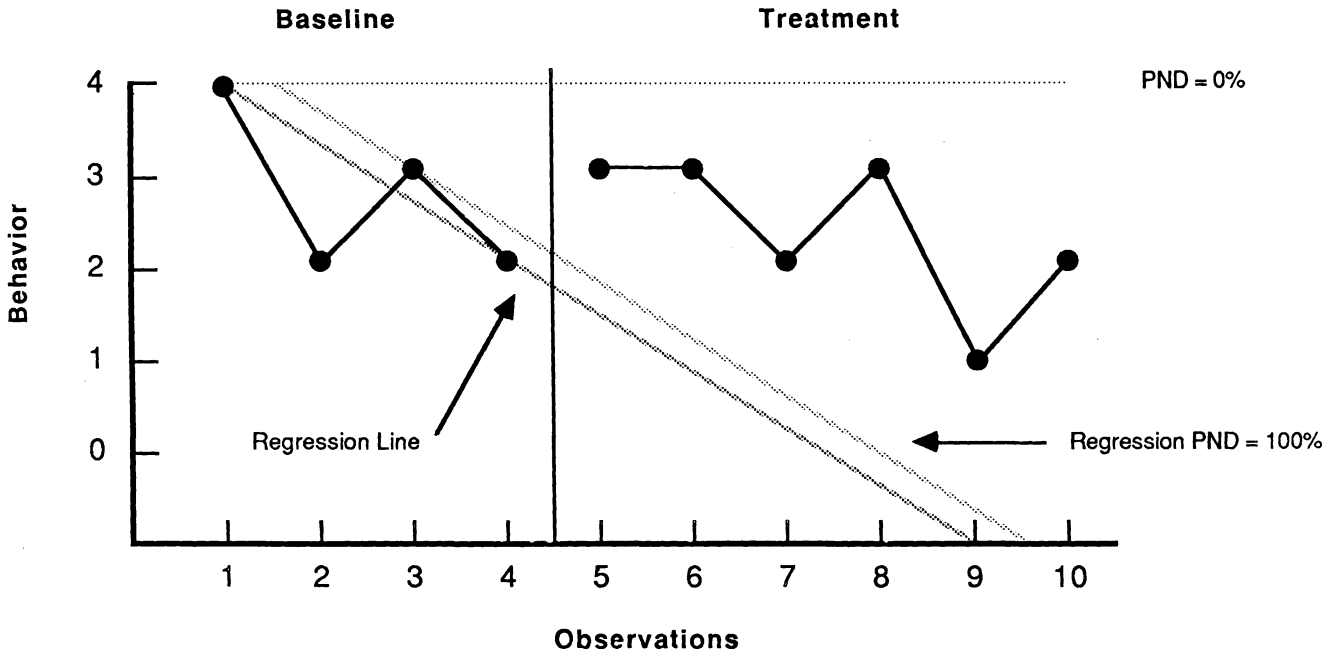

Figure 1. Data displaying idiosyncratic effects of baseline trend for PND calculation.

tive criteria for this judgment were lacking. What we employed was a reliability procedure in which different raters initially agreed on examples of inappropriate baseline trends and later established reliability of such judgments. Although this procedure was highly reliable for our own purposes, in practice we encountered very few examples of such problems. Perhaps future researchers will find a need to further refine these (or any other) conventions, but according to data reported by Huitema, such baseline trends may not be at all common. In fact, it is our impression that editors of behavioral journals actively discourage the publication of research studies with unstable baseline trends.

Summary

We appreciate the opportunity to respond to Dr. White's comments. We regard them as thoughtful and constructive, although we do not believe he has either discredited our PND metric or recommended a more useful alternative. Our own view is that "the proof of the pudding is in the eating," and we refer the reader to our published synthesis efforts. We think that our results are basically in agreement with the conclusions of the original research reports, and that a meaningful outcome measure is provided by which different research reports can conscionably be objectively compared. Of course, it is necessary to temper such procedures with "common sense," and, in fact, we would never recommend the coding of a PND outcome if it flew in the face of visual inspection procedures. As efforts continue in this area, we hope procedures we have recommended will be modified with respect to actual, rather than hypothesized, problems.

We agree with Dr. White that further application of this, and possible alternative methods, are needed. If alternative procedures that produce more systematic, objective results are found, we welcome them. Until such methods are forthcoming, we stand by the PND outcome as objective, interpretable, and meaningful.

Thomas E. Scruggs, *PhD, is a visiting assistant professor in the Department of Education, Special Education Section, and Director of the Purdue Achievement Center, Purdue University, West Lafayette, IN. He received a PhD from Arizona State University in 1982. His research interests include assessment, learning and memory, and research synthesis.* **Margo A. Mastropieri,** *PhD, is an assistant professor in the Department of Education, Special Education Section, Purdue University. She received a PhD from Arizona State University in 1983. Her research interests include teacher effectiveness, mnemonic strategies, prose comprehension, and research synthesis.* **Glendon Casto,** *PhD, is the Director of the Early Intervention Research Institute; acting director of the Developmental Center for Handicapped Persons; and professor of psychology, Utah State University, Logan. He received his PhD from the University of Utah. His research interests include cognitive development, assessment in early intervention, early intervention programmatic and efficacy research, and research synthesis.*

Authors' Note

1. The work reported in this article was carried out in part with funds from the U.S. Department of Education (Contract No. 300-82-0367) to the Early Intervention Research Institute at Utah State University.

(*continued on p. 39*)