The Behavior Analyst 2000, 23, 101–106 No. 1 (Spring)

In Response

Quantitative Summaries of Single-Subject Studies: What Do Group Comparisons Tell Us About Individual Performances?

Alan Baron and Adam Derenne
University of Wisconsin–Milwaukee

Kollins, Newland, and Critchfield (1999) responded to our comments about their review by arguing that their quantitative summary was not a meta-analysis and should not be criticized in these terms. We reply that regardless of what they call their review, it included confounding effects that make interpretations of the results problematic. Kollins et al. also argued that unexpected findings of the sort they reported can serve as a spur for further research. We reply that the understanding of findings that deviate from existing knowledge may well require empirical investigation. Such endeavors, however, should begin with an evaluation of the review procedures that suggested the existence of the differences. Finally, we emphasize that quantitative summaries of individual data are, in the end, a form of group comparison. The implications of using group methods to clarify individual data deserve frank recognition in discussions of the outcomes.

Kollins, Newland, and Critchfield's (1997) quantitative review of the literature on human choice attracted our attention for several reasons. Foremost is that efforts to summarize the complex human operant literature are interesting and important. In this case, the summary yielded some unusual outcomes. For example, the report suggested that humans are more sensitive to contingencies in naturalistic settings than in laboratory settings, a finding that some may take as a call to abandon laboratory research. In addition, their analytic method had many features of a meta-analysis (a quantitative "analysis of analyses"). Although meta-analysis as a way of summarizing the literature is increasingly used, the procedures have engendered considerable controversy. Additional problems emerge when the method is extended to single-subject data. This conflux of events—a set of unusual results obtained with an unusual analytic method, plus the fact that we discovered the review in a journal not ordinarily concerned with operant psychology (*Psychonomic Bulletin and Review*)—suggested the value of an article that would air the issues before the behavior-analytic community (see Derenne & Baron, 1999).

In light of Kollins, Newland, and Critchfield's (1999) reply to our comments, perhaps it would be well for us to reiterate that we see considerable value in their effort to make sense of the puzzling human operant literature. For better or worse, we have contributed to that literature from its early stages of development (e.g., Antonitis, Frey, & Baron, 1964; Baron & Kaufman, 1966). We also joined with such figures as Don Hake (1982) and Harold Weiner (1970) in calling for increased use of human subjects in the experimental study of basic processes (e.g., Baron & Galizio, 1983; Baron & Perone, 1982). As illustrated by Kollins et al.'s publication, this is coming to pass. Our views about the importance of research with humans led us to commend

We thank Marshall Dermer for his helpful comments.

Please address correspondence to either author at the Department of Psychology, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201 (E-mail: ab@uwm.edu or aderenne@uwm.edu).

Kollins et al. for undertaking their review, and we complimented them for their "numerous insights and cogent comments" (Derenne & Baron, 1999, p. 39). Perhaps we were remiss in not also complimenting them for their original method of summarizing the findings in the form of Q-Q plots, and we take this opportunity to do so.

We were disappointed, however, that Kollins et al.'s (1999) response did not engage concerns that we regard as worthy of discussion. Instead, they broadly characterized our views as inaccurate because we labeled their effort a meta-analysis, and irrelevant because our concerns were appropriate for analyses of group-statistical rather than single-subject studies. Our further comments center on these matters.

The critical issue is not the label that should be applied to Kollins et al.'s (1997) review. They do concede that their effort was a "form of meta-analysis" in that "data were synthesized across studies and a quantitative description of the relations among such data was used" (Kollins et al., 1999, pp. 152, 153). But they make the further point that unlike the usual meta-analytic study, they were not concerned with the average performance of groups, and they did not apply the methods of inferential statistics to calculate measures of effect size. We do not see much profit in debating whether it is better to call a review that does not include these latter features a "meta-analysis" or a "form of meta-analysis." More important is the recognition that there are logical pitfalls awaiting researchers who combine data from different studies. Implicit in our comments was the caution that these pitfalls are similar in a number of respects for those who work at the level of the individual subject and those who use group-statistical methods.

An instructive parallel can be found in traditional views of the way behavior analysts conduct experiments. As we know, the behavior-analytic approach to research is regarded with considerable suspicion in some circles (see Dermer & Hoch, 1999, for a review of textbooks on method). Single-subject experiments are said to be "quasi-experiments" (something less than an experiment) because the analysis does not aggregate data from different subjects or use the methods of inferential statistics to evaluate the outcomes. Over the years, behavior analysts have objected to this characterization. By history and tradition (e.g., Sidman, 1960; Skinner, 1938), we regard experimental design as a conceptual, not a statistical, matter. Those features associated with conventional experiments—the averaging of group data and statistical tests of the results—are best regarded as appendages to the experimental method, and they can be dispensed with entirely when experiments are designed in ways that employ experimental control to isolate variables of interest (Baron & Perone, 1998). Bachrach (1981) expressed this view when he wrote, "Research is *not* statistics" (p. 1).

Similar distinctions can be drawn for techniques that combine data across studies. The logic of the comparisons stands independently of whether inferential statistics are used to buttress confidence in the interpretations. The four logical concerns expressed in our commentary are not original. They have been discussed in connection with single-subject research (e.g., Salzberg, Strain, & Baer, 1987) as well as the traditional literature (e.g., Streiner, 1991). In the interests of providing direction for future literature summaries, we will reiterate them here. The common theme is that of *confounding*—the confusion of a correlated variable with the one believed critical for a particular outcome. In the case of the individual experiment, confounding lays open to question the effects of the controlling variable. The hazards in the case of meta-analysis are much the same.

First, in their effort to determine whether human performances are more variable than those of the pigeons and rats studied by Baum (1979), Kollins et al. (1997) exaggerated species dif-

ferences by not taking into account variables known to influence choice. They correctly noted that procedures in the human laboratory are not well standardized (cf. Perone, Galizio, & Baron, 1988), and they focused on several procedural differences, such as whether the reinforcer was monetary or nonmonetary. However, in so doing they included conditions that were excluded in Baum's analysis. For example, Kollins et al.'s data pool contains studies that used single as well as concurrent schedules, and studies that used concurrent variable-ratio schedules as well as concurrent variable-interval schedules. They also pooled time-matching and response-matching data (Baum kept these measures separate), and they disregarded the developmental levels of the subjects. These deviations from Baum's procedures undermine conclusions about the extent or origin of human–nonhuman differences.

Second, Kollins et al. (1997) did not evaluate differences in the rigor of the studies that were entered into the data pool. Sidman (1960) has pointed out that equal confidence cannot be placed in published research because of differences in control procedures. Most notably, the reliability of data from individual subjects is compromised when findings are not based on stable performances. Nevertheless, Kollins et al.'s analysis included data without regard to stability, with the consequence that the analysis confounded the reliability of the individual studies with variations in the overall pattern of results.

Third, Kollins et al. (1997) entered data into the analysis regardless of the number of subjects in a particular experiment. An implicit assumption of meta-analysis is that data from the subjects in each experiment contribute equally to the outcomes. However, in Kollins et al.'s analysis of the laboratory experiments, more than half the data points were taken from a single experiment, and 7 of the 17 reports came from the same laboratory. Their method of selecting data for their analysis opened the door to confounding effects between effects of the variable of interest and the idiosyncratic procedures of a given researcher or laboratory.

Fourth, identification of specific procedures (so-called moderator variables) that may contribute to overall variation requires that each moderator variable operates independently. This principle was not adhered to in Kollins et al.'s (1997) analysis. For example, they present data that appear to show that sensitivity was reduced when reinforcement depended on button pressing rather than some other response (one of the moderator variables). However, virtually all of the data points from the button-pressing studies also used procedures in which money, rather than some other event, was the reinforcer (another of the moderator variables). Similarly, consideration of the individual experiments indicates that the subjects in the laboratory studies were adults, whereas those in the naturalistic studies usually were children or adolescents. These intercorrelations confound efforts to determine which variable or variables produced a particular outcome.

Kollins et al. (1999) responded to these criticisms by reminding us that they had indeed expressed qualifications about the results, as indicated by the following quotation from their article: "Because most studies differ procedurally in several ways, the binary comparisons conducted here are artificially simple in that they do not reflect the possible interactions among factors that could influence sensitivity to reinforcement. . . . Many other kinds of comparisons are possible, as the relevant studies differ procedurally on many dimensions" (Kollins et al., 1999, p. 154). We stand corrected about our blanket statement that "qualifications were not expressed in the article" (Derenne & Baron, 1999, p. 38). The fact remains, however, that although Kollins et al. acknowledged in general the limitations of their analysis, they failed to address the specific is-

sues that we raised. The possibility of misunderstanding these specific issues prompted us to write our commentary. In our view, each of the questionable findings required discussion of alternative interpretations. Thus, we observed that differences with regard to the type of response can be equally attributed to the type of reinforcer and that differences between laboratory and naturalistic studies might have something to do with the developmental level of subjects.

Kollins et al. (1999) also offered us advice about the way to view "unexpected findings." They noted in the concluding paragraphs of their reply that their results "identified many testable hypotheses about the role of procedural variations on human choice" (Kollins et al., p. 155). Unexpected results do deserve attention, and that is why we commented on Kollins et al.'s analysis. Our differences with them revolve around the sorts of hypotheses that should be pursued in attempting to account for such findings.

At least four distinctly different sets of circumstances can contribute to discrepant findings, that is, results that are inconsistent with other results, or results that deviate from "theory or common observation" (Kollins et al., 1999, p. 154), which, after all, are summaries of empirical findings. The first set is the one emphasized by Kollins et al. (1997, 1999). Replications may yield different results because the replication failed to include essential variables. For example, the inability to replicate animal findings in the human laboratory may stem from the use of reinforcers of lesser potency or the neglect of important setting events. A second source of variation originates in the care with which the experiment was conducted. For example, background variables may be poorly controlled, as when, in a study of fixed-interval reinforcement in humans, the researcher fails to check whether the subject has brought a watch into the experimental room. Third, different results may reflect failure to adhere to the data-analytic procedures required by single-subject designs, in particular, attainment of stability. Use of more or less rigorous stability criteria in different studies reduces the likelihood that results will be similar. Also, variations in the results of different studies may be a consequence of deficiencies in the summary procedures themselves. The logical problems we have mentioned fall in this category.

In their reply, Kollins et al. (1999) emphasized that "methods matter" (p. 149), and they recommended further investigation of the differences in method that they detected. In so doing, they focused only on the first of the four sources of variation enumerated above. They did not consider the second and third sources of variation, that is, variation resulting from grouping experiments without regard to the level of care with which the research was conducted or the extent to which the usual requirements of a single-subject data analysis were met. They also did not consider the fourth source: the extent to which their own methods of summarizing the literature may have clouded identification of general principles that cut across the different studies.

Finally, we continue to find an incongruity in efforts to conduct quantitative summaries of single-subject studies. Whatever the arguments for adopting a single-subject approach to the analysis of behavior, these same arguments apply to summaries of the results of different studies. Unavoidably, quantitative reviews aggregate the performances of individual subjects, and thus contain the essential features of group comparisons. Both group-statistical approaches to research and quantitative summaries of different studies accept variation as an inherent feature of the results (either from subject to subject or from experiment to experiment). Moreover, as we noted in our commentary, such groupings in a meta-analysis are open to further question because commonalities are accomplished by selection rather than by the

experimental manipulation that is the hallmark of behavior-analytic studies. In other words, the meta-analyst selects studies that exemplify the variable, as when Kollins et al. (1997) developed groupings of laboratory versus natural-setting studies. The inevitable consequence is that a study selected on the basis of some distinguishing characteristic carries along with it the baggage of a host of other variables, both known and unknown.

To bring home the point that meta-analyses of single-subject data embody many of the features of traditional group-statistical procedures, imagine a researcher interested in whether the setting—laboratory or natural—influences the way subjects perform. He or she might observe groups of subjects within each of these environments, and then go on to analyze the results using the same Q-Q plot method employed by Kollins et al. (1997). It seems evident that the merit of our thought experiment would depend on the extent to which the researcher had dealt with potential confounding variables. We would hope that the subjects were similar in age and other individual characteristics, that procedural details were the same except for the variable under investigation (the setting), and that the between-groups results were subjected to some sort of statistical test (unless differences were unambiguous through inspection). Indeed, anything less would be considered a serious failing of the research.

Moreover, some might go further and object to the between-groups design of the experiment, with its neglect of the variables that have produced the variation from individual to individual within the same condition. In this regard, Q-Q plots are not a remedy. Although this way of presenting the data may express the individual variation (as do frequency distributions), Q-Q plots in no way account for it. We would not go so far as to argue that between-groups comparisons should be excluded from the methods of behavior analysis (cf. Sidman, 1960). Sometimes they are demanded by the issue under investigation, as when age differences are the object of study (Baron, 1990). However, we do assert that comparisons should be arranged in ways that minimize the confounding of variables. This is the bone of our contention with Kollins et al. (1997).

In summary, we hope that the present discussion is a constructive step toward developing the "best practice" that Kollins et al. (1999, p. 153) believe is lacking. We acknowledged in our earlier comments that the alternate way of summarizing the literature—the traditional narrative literature review—has problems of its own. As many have pointed out (e.g., Glass, McGaw, & Smith, 1981), conclusions rely heavily on the reviewer's subjective evaluations of the literature, and such judgments may differ markedly from review to review. However, the redeeming value of the traditional review may be that it requires the reviewer to carefully scrutinize each of the relevant studies, and allows him or her to give different studies different weight depending on nuances of the procedures and subtle aspects of the results.

The question of how to go about summarizing the single-subject literature remains unresolved. By comparison with the traditional qualitative review, Kollins et al.'s (1997) approach has the merit of achieving quantification of the summary process. The potential advantages from the standpoint of precision and objectivity are obvious. The method suffers, however, because efforts to combine single-subject results from numerous studies in terms of a common metric divert the analyst from a close consideration of the procedures and data that characterize the individual subjects in each of the studies (Salzberg et al., 1987). In this regard, the meta-analyst may be said to neglect the individual study just as group analysts neglect the individual subject. Perhaps the most promising approach at this time is one that melds the best features of quantitative and qualitative methods of literature re-

view. When data from different studies are grouped for the purpose of a quantitative analysis, care must be given to the rules for including a particular study within a particular category. In addition, frank recognition must be given to the fact that the procedure has many of the characteristics of the traditional experimental method that behavior analysts have deplored (in particular, those of the so-called ex post facto experiment). Under the circumstances, it seems unavoidable that some of the techniques of traditional methods must be adopted. For example, if comparisons are made on the basis of selection rather than manipulation of variables, then special emphasis must be placed on matching procedures; and, if there is substantial variation within the data from the resulting groupings, then statistical procedures may be needed to evaluate the reality of any between-groups differences.

REFERENCES

Antonitis, J. J., Frey, R. B., & Baron, A. (1964). Group operant behavior: Effects of tape-recorded verbal reinforcers on the bar-pressing behavior of preschool children in a real-life situation. *Journal of Genetic Psychology, 105,* 311–331.

Bachrach, A. J. (1981). *Psychological research: An introduction* (4th ed.). New York: Random House.

Baron, A. (1990). Experimental designs. *The Behavior Analyst, 13,* 167–171.

Baron, A., & Galizio, M. (1983). Instructional control of human operant behavior. *The Psychological Record, 33,* 495–520.

Baron, A., & Kaufman, A. (1966). Human, free-operant avoidance of "time out" from monetary reinforcement. *Journal of the Experimental Analysis of Behavior, 9,* 557–565.

Baron, A., & Perone, M. (1982). The place of the human subject in the operant laboratory. *The Behavior Analyst, 5,* 143–158.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 45–91). New York: Plenum.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior, 32,* 269–281.

Derenne, A., & Baron, A. (1999). Human sensitivity to reinforcement: A comment on Kollins, Newland, and Critchfield's (1997) quantitative literature review. *The Behavior Analyst, 22,* 35–41.

Dermer, M. L., & Hoch, T. A. (1999). Improving descriptions of single-subject experiments in research texts written for undergraduates. *The Psychological Record, 49,* 49–66.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Hake, D. F. (1982). The basic-applied continuum and the possible evolution of human operant social and verbal research. *The Behavior Analyst, 5,* 21–28.

Kollins, S. H., Newland, M. C., & Critchfield, T. S. (1997). Human sensitivity to reinforcement in operant choice: How much do consequences matter? *Psychonomic Bulletin & Review, 4,* 208–220. Erratum: *Psychonomic Bulletin & Review, 4,* 431.

Kollins, S. H., Newland, M. C., & Critchfield, T. S. (1999). Quantitative integration of single-subject studies: Methods and misinterpretations. *The Behavior Analyst, 22,* 149–157.

Perone, M., Galizio, M., & Baron, A. (1988). The relevance of animal-based principles in the laboratory study of human operant conditioning. In G. Davey & C. Cullen (Eds.), *Human operant conditioning and behavior modification* (pp. 59–85). New York: Wiley.

Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *Remedial and Special Education, 8,* 43–48.

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

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century.

Streiner, D. L. (1991). Using meta-analysis in psychiatric research. *Canadian Journal of Psychiatry, 36,* 357–361.

Weiner, H. (1970). Human behavioral persistence. *The Psychological Record, 20,* 445–456.