In Response

Human Sensitivity to Reinforcement: A Comment on Kollins, Newland, and Critchfield's (1997) Quantitative Literature Review

Adam Derenne and Alan Baron
University of Wisconsin–Milwaukee

In a quantitative review of human operant experiments, Kollins, Newland, and Critchfield (1997) found that humans are less sensitive to reinforcement contingencies than nonhumans are. Human performances were not as consistent with the matching law, and they were more variable from subject to subject. Some of the variables correlated with reduced human sensitivity were surprising. These included collection of the data under more controlled conditions (laboratory rather than naturalistic settings), and inclusions of discriminative stimuli correlated with alternative sources of reinforcement. We discuss these unexpected findings in the light of criticisms that have been leveled against meta-analytic literature reviews (e.g., the wisdom of grouping studies with widely diverse methods), and we suggest ways of improving future analyses of the behavior-analytic literature.

Key words: humans, sensitivity, matching, concurrent schedules, choice, meta-analysis

Sensitivity to reinforcement can be measured by assessing the extent to which subjects match ratios of response and reinforcement rates. In a recently published article, Kollins, Newland, and Critchfield (1997) reviewed the literature on human sensitivity and compared the results to those reported for nonhumans. An unusual feature of their review was the application of meta-analytic procedures, usually associated with group-statistical research methods, to a literature that has employed single-subject methods. Also noteworthy were some of their surprising findings. For example, human responding was found to be less sensitive when the research was conducted in the laboratory rather than in naturalistic settings (i.e., under conditions of

lesser control) and when discriminative stimuli were involved (i.e., under conditions that increased stimulus differences between the alternate sources of reinforcement). In this article, we comment on some potential problems with their analysis, first from the standpoint of general criticisms leveled against meta-analysis, and then with regard to the wisdom of melding single-subject and meta-analytic methods.

Kollins et al.'s (1997) analysis was conducted in two phases. The first part used a previous review by Baum (1979) as a benchmark. Baum summarized experiments in which nonhuman subjects (almost all were pigeons) responded on concurrent variable-interval schedules of reinforcement, and he found a reasonably close correspondence between response and reinforcement ratios. The extent of this correspondence was quantified by the slope of the matching function: A slope of 1 indicates perfect matching, whereas slopes less than 1 (undermatching) or greater than 1 (overmatching) indicate lesser degrees of sensitivity. The median slope of the functions aggregated from the subjects in the different ex-

This article is based on a report given by the first author when he was a student in the second author's course, "Seminar in Operant Behavior." We thank Marshall Dermer for his helpful comments on an earlier version.

Requests for reprints and correspondence concerning this article can be addressed to either Adam Derenne or Alan Baron at the Department of Psychology, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201 (E-mail: aderenne@uwm.edu or ab@uwm.edu).

periments was 0.85, indicating that nonhumans showed a moderate tendency to undermatch. By comparison, in Kollins et al.'s replication of Baum's analysis with data drawn from the human literature (they used data from both single-response and concurrent studies), the median slope was 0.70. Thus, the performances of the human subjects were characterized by lesser degrees of sensitivity than the nonhumans, as indicated by greater undermatching. Also striking were differences in the variability of the slope distributions. As indexed by the interquartile ranges, the human subjects were considerably more variable (humans, 0.31 to 1.01; nonhumans, 0.75 to 0.97).

The second part of Kollins et al.'s analysis involved a series of subanalyses of the human data. Their goal was to identify variables that might contribute to human insensitivity and variability. In these subanalyses, they found that human sensitivity was lower and subject-to-subject variation was greater, (a) when the operant response was a simple button press, (b) when the consequences of responding involved monetary reinforcement, (c) when the experiment was conducted in the laboratory, and (d) when discriminative stimuli were correlated with the two sources of reinforcement.

Kollins et al. are to be commended for their effort to clarify the complicated human operant literature. Their findings bear careful scrutiny, however, insofar as they run counter to both theory and common observation. As we noted already, it seems remarkable that naturalistic studies would yield more consistent results than those conducted within controlled laboratory environments, or that sensitivity should be reduced by adding discriminative stimuli. In addition, meta-analysis as a way of synthesizing results from different studies is a controversial procedure, and the problems are compounded when the method is applied to results from single-subject studies (Salzberg, Strain, & Baer, 1987). Although the techniques that have been labeled

meta-analysis are diverse, the distinguishing feature of the method is that a common measure is derived from the results of the studies in question (in the present case, the slope values). These values then are statistically combined in an effort to discern a uniform pattern across the studies. (As one text put it, "in essence, meta-analysis treats each study as if it were a subject in a study of studies," Stern & Kalof, 1996, p. 252.) Meta-analyses of the behavioral and biomedical science literatures almost invariably use group averages from the studies to provide the separate data points. By comparison, Kollins et al.'s meta-analysis followed the unusual procedure of using data from the individual subjects within each of the studies. We will discuss the implications of this difference in subsequent sections of this article.

These considerations—the unusual results and the analytic method—suggest that it would be useful to address the results of the review in the context of some of the systemic problems that have plagued meta-analyses in general (Eysenck, 1994; Oakes, 1986; Salzberg et al., 1987; Streiner, 1991). These problems have been discussed within the meta-analytic literature in terms of four broad issues.

Criteria for Including or Excluding Studies

It is apparent that the methods of the experiments in Kollins et al.'s pool of human data varied widely, including differences in research setting, type of reinforcer, and topography of the operant response. A controversial issue in meta-analysis is the extent of allowable variation from study to study (Streiner, 1991). For example, meta-analyses of the effectiveness of psychotherapy have been questioned because they included studies in which clients were suffering from a variety of disorders and were treated with a variety of therapies. Critics have argued that such broad variations in method cloud the magnitude of any general effects be-

cause the data pool includes ineffective as well as effective techniques (Eysenck, 1994; Oakes, 1986). An alternative view is that grouping of diverse studies is quite appropriate when a general issue is being addressed, but it is an error to end the analysis at this point. What also is required are subanalyses designed to isolate the specific variables on which the general outcome depends (Streiner, 1991). Moreover, the analysis should specify clearly the boundaries of the data pool. One authority put it this way:

The point here is not that there is a "right answer" as to which studies should be included or excluded, but that each meta-analyst should state clearly at the onset the criteria for excluding some of the studies identified, and then explain why the criteria are appropriate for the specific meta-analysis. (Cook et al., 1992, p. 297)

In their review, Kollins et al. correctly followed the two-step procedure. They first compared the human experiments, which employed a range of different procedures, to the nonhuman experiments analyzed by Baum (the outcome is shown in Kollins et al., Figure 3). They then examined specific variables that might have contributed to the differences (see Kollins et al., Figure 4). Notably absent in their report, however, is information about the criteria used for including or excluding studies.

The particular studies that were selected introduced potential sources of variation of which we will mention four: First, although the analysis is titled as being concerned with "operant choice," the review included experiments with single schedules (e.g., Bradshaw, Szabadi, & Bevan, 1979) as well as concurrent schedules. Second, although the conventional procedure for studying matching involves choices between interval schedules, the analysis included an experiment in which concurrent variable-ratio schedules were used (Mace, McCurdy, & Quigley, 1990). Third, among the group of studies classified as "naturalistic," some involved problem individuals in educational settings (e.g., Martens &

Houk, 1989) and others involved college students in the laboratory (e.g., Beardsley & McDowell, 1992). Fourth, although Baum reported different outcomes when time-allocation rather than response-allocation measures were used, Kollins et al.'s analysis pooled results across these two measures (e.g., Takahasi & Iwamoto, 1986).

Kollins et al.'s liberal selection criteria may have a bearing on their finding that human performances are more variable. Although it is possible that such variation is an inherent aspect of human performance, the differences between humans and nonhumans more likely reflect the greater variation in method (subjects, apparatus, procedures) when operant research is conducted with humans (Perone, Galizio, & Baron, 1988). By comparison with the varied procedures within the pool of human experiments, the nonhuman experiments reviewed by Baum (1979) involved a set of restricted procedures. Almost all of the data (93 of 99 data points) came from experiments in which pigeons pecked keys in standard operant conditioning chambers to receive food reinforcement delivered on concurrent variable-interval schedules. Thus, the greater variation of the data from the studies with humans may be as much due to the reviewers' selection procedures as to inherent features of human operant performance.

Differences in the Rigor of the Studies

It is inevitable that the studies grouped in a meta-analysis will vary not only in terms of their procedures but also in their rigor. Within operant psychology, there is agreement about the criteria for sound methodology in steady-state research with human and nonhuman subjects (Lattal & Perone, 1998; Sidman, 1960). We will not attempt to consider all of these criteria. One that stands out is the extent to which the procedure involves sequential observations of individual performances and steady-state data. The hu-

man experiments in Kollins et al.'s data pool varied considerably in this regard. For example, Conger and Killeen (1974) synthesized a single matching function from observations of 5 subjects that lasted only 15 min. By comparison, each slope value in Bradshaw, Szabadi, and Bevan's (1977) experiment was derived from the performance of an individual subject using data collected over a period of 30 days or more.

As with the question of excluding studies because they are too diverse, there are different views on the wisdom of excluding studies on methodological grounds. A common criticism is that inclusion of flawed studies within a meta-analysis will generate flawed conclusions. Defenders of liberal criteria have responded that identifying methodological soundness is difficult at best (Cook et al., 1992; Glass, McGaw, & Smith, 1981). A compromise is to include all studies regardless of quality, but then to designate methodological adequacy as one of the variables in the analysis (Streiner, 1991). However, Kollins et al. did not take this step. For example, they might have classified the experiments in terms of their stability criteria to determine whether experiments with more stringent criteria yielded higher levels of sensitivity (slopes closer to 1). In the absence of such analyses, their review leaves open the possibility that variations among the human experiments reflected methodological differences rather than true differences in sensitivity.

Differences in the Number of Outcomes per Study

If a small number of studies contribute disproportionately to the data pool, meta-analytic outcomes may be unduly influenced by idiosyncratic features of those studies (Streiner, 1991). A source of concern in Kollins et al.'s analysis is that one experiment (Takahasi & Iwamoto, 1986) contributed more than half of the data points for the group of

laboratory studies, and 7 of the 17 laboratory reports came from the same laboratory (Kollins et al., Table 1). These imbalances in the general data pool were accompanied by imbalances within the specific subanalyses (summarized in Kollins et al., Figure 4). For example, in the comparison of the "button press" category with the "other response" category, values for the "button press" slopes involved 245 data points from 12 studies, whereas values for the "other response" category totaled only 34 data points from five studies (Kollins et al., Figure 4B).

The exact contribution of such imbalances to the outcomes of the analyses is difficult to assess. Compounding the problem is that performances within the subanalyses were less variable and more sensitive for the distribution with the fewer data points. This pattern may be seen, for example, in differences in slope values as a function of schedule-correlated stimuli (Kollins et al., Figure 4C). Sensitivity was greater when stimuli were absent, a finding that the authors noted was "the opposite of what might be expected" (p. 213). We cannot offer a simple account of these correspondences. We do see a need to qualify the findings in the light of the disproportionate contribution of some studies to the data pool and the apparent relation between outcomes and number of cases.

Independence of the Moderator Variables

We noted the need for subanalyses to identify which procedures (moderator variables) contributed to the overall outcome (Streiner, 1991). Because qualifications were not expressed in the article, the implication is that each of the variables that influenced human sensitivity operated independently. Closer examination of the procedures used to classify the studies reveals that this was not the case.

A case in point is the comparison of "points for money" with "other rein-

forcers" (Kollins et al., Figure 4B). In 11 of the 13 studies in which monetary reinforcers were used, the response was button pressing, whereas in three of the four studies with nonmonetary reinforcers, a different type of response was used (Kollins et al., Table 1). These correspondences make it impossible to determine if the lower level of sensitivity when subjects press buttons for money is due to the type of response, the type of reinforcer, or some sort of interaction between the two. These same variables were confounded in the comparison of "laboratory" and "naturalistic" studies (Kollins et al., Figure 4D). Monetary reinforcement was rarely used in the naturalistic studies (one of nine studies), and all of the naturalistic studies employed some response other than button pressing (Kollins et al., Table 2). To complicate matters further, the subjects in the naturalistic studies usually were children or adolescents, whereas the subjects in the laboratory experiments were adults.

Intercorrelations between moderator variables are a well-known problem in meta-analytic studies, and potential biases are usually corrected through inferential statistical procedures such as multiple regression. However, this was not attempted by Kollins et al. Moreover, unlike Baum (1979), who also had to cope with variability in the data, Kollins et al. did not support any of their conclusions through statistical tests. Because meta-analysis is, essentially, a group procedure (e.g., as when a group of laboratory studies is compared to a group of naturalistic studies), tests of significance are required. These omissions call for qualifications of conclusions about the variables that may contribute to human sensitivity.

Conclusion

Despite the questions we have raised, we give Kollins et al. (1997) considerable credit for their effort to confront some puzzling outcomes when humans serve as subjects in operant research. Moreover, their inter-

pretations include numerous insights and cogent comments about human performances. However, their approach poses fundamental conceptual questions because it melds data from the behavior-analytic literature (single-subject experiments) with procedures more commonly associated with group-statistical research designs. The logic of meta-analysis is to treat the results of each experiment as if they constituted a single subject, to aggregate these values, and then to employ conventional inferential statistical methods to determine if average differences are reliable. By comparison, behavior-analytic research strategies eschew the aggregation of data from different subjects and the use of inferential statistics in favor of methods that emphasize repeated observations of the same subject under controlled conditions.

As noted by Salzberg et al. (1987), such a mixing of meta-analytic procedures and single-subject results strikes a discordant note. By history and tradition, the experimental analysis of behavior places heavy reliance on controlled observations. The experimenter manipulates variables, makes observations of steady states, and reaches conclusions about the controlling influences of the variables. By comparison, meta-analytic procedures are more closely allied with correlational methods. Unlike controlled experimentation, the variables of a meta-analysis are arrived at by selection rather than by manipulation: The meta-analyst must seek them out from studies already reported in the literature. The obvious limitation of an approach that relies on selection is that possible actions of the focal variables will be confounded with other variables, unknown as well as known. Conclusions about the controlling influence of variables can only suffer as a result. This is the origin of many of the concerns we have expressed.

Although we have misgivings about meta-analytic reviews of single-subject experiments, we are not prepared to advocate that the approach be aban-

done. As discussed by proponents of meta-analysis, the alternative—the traditional discursive literature review—has its own peculiar limitations and biases (e.g., Cook et al., 1992; Glass et al., 1981). Instead, our modest proposal is simply that the same care that has traditionally characterized behavior-analytic research in the laboratory be extended to the procedures of quantitative literature reviews. In the interests of objectivity, meta-analyses often involve the uncritical grouping of experiments, with the main requirement that they all be directed toward a common issue. More in the spirit of an experimental analysis are procedures that closely scrutinize each study to determine whether it should be included within the general data pool, and, if admitted, to which category it should be assigned.

Indeed, Kollins et al. advocated one version of such a selection strategy in their discussion of some general ways to strengthen human operant research. They suggested that human operant experiments often neglect the potential role of establishing operations. In the animal laboratory, establishing operations can be experimentally manipulated, as when nonhuman subjects are deprived of food or water to establish these events as reinforcers. Although parallel manipulations usually are impractical in human research, an alternative proposed by Kollins et al. is to select subjects in terms of existing levels of the operation. For example, when money is used as the reinforcer, subject-to-subject variability might be reduced by screening subjects in terms of their financial circumstances. Similar steps can be envisioned for a range of other human variables that cannot be controlled experimentally.

In their discussion of establishing operations, Kollins et al. were expressing an important principle of effective behavioral research: the need to carefully select the characteristics of subjects who will participate in the experiment. It is not a large step to extend this notion to the meta-analysis, that is,

to procedures that employ the individual study rather than the individual subject as the unit of analysis. Selection of a study for inclusion within a meta-analysis means that the levels of a host of establishing operations also have been selected, such as the type of reinforcer, characteristics of the subjects, and whether the study is conducted in the laboratory or the clinic. As pointed out by Kollins et al., such variables, if not properly addressed, can obscure the results and cloud interpretations of individual experiments. Our purpose here is to emphasize that the same considerations pertain to the reanalysis of the results of already-completed experiments.

REFERENCES

- Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior*, 32, 269–281.
- Beardsley, S. D., & McDowell, J. J. (1992). Application of Herrnstein's hyperbola to time allocation or naturalistic human behavior maintained by naturalistic social reinforcement. *Journal of the Experimental Analysis of Behavior*, 57, 177–185.
- Bradshaw, C. M., Szabadi, E., & Bevan, P. (1977). Effect of punishment on human variable-interval performance. *Journal of the Experimental Analysis of Behavior*, 27, 275–279.
- Bradshaw, C. M., Szabadi, E., & Bevan, P. (1979). The effect of punishment on free-operant behavior in humans. *Journal of the Experimental Analysis of Behavior*, 31, 71–81.
- Conger, R., & Killeen, P. (1974). Use of concurrent operants in small group research. *Pacific Sociological Review*, 17, 399–416.
- Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation*. New York: Russell Sage Foundation.
- Eysenck, H. J. (1994). Meta-analysis and its problems. *British Medical Journal*, 309, 789–792.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Kollins, S. H., Newland, M. C., & Critchfield, T. S. (1997). Human sensitivity to reinforcement in operant choice: How much do consequences matter? *Psychonomic Bulletin and Review*, 4, 208–220. Erratum: *Psychonomic Bulletin and Review*, 4, 431.
- Lattal, K. A., & Perone, M. (Eds.). (1998). *Handbook of research methods in human operant behavior*. New York: Plenum.

- Mace, F. C., McCurdy, B., & Quigley, E. A. (1990). A collateral effect of reward predicted by matching theory. *Journal of Applied Behavior Analysis*, 23, 187-205.
- Martens, B. K., & Houk, J. L. (1989). The application of Herrnstein's law of effect to disruptive and on-task behavior of a retarded adolescent girl. *Journal of the Experimental Analysis of Behavior*, 51, 17-26.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York: Wiley.
- 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.
- Stern, P. C., & Kalof, L. (1996). *Evaluating social science research, second edition*. New York: Oxford University Press.
- Streiner, D. L. (1991). Using meta-analysis in psychiatric research. *Canadian Journal of Psychiatry*, 36, 357-361.
- Takahasi, M., & Iwamoto, T. (1986). Human concurrent performances: The effects of experience, instructions, and schedule correlated stimuli. *Journal of the Experimental Analysis of Behavior*, 45, 257-267.