

The Good, the Bad, and the Aggregate

Thomas S. Critchfield
Illinois State University

M. Christopher Newland
Auburn University

Scott H. Kollins
Western Michigan University

To evaluate progress and focus goals, scientific disciplines need to identify relations that are robust across many situations. One approach is the literature review, which characterizes generality across studies. Some writers (e.g., Baron & Derenne, 2000) claim that quantitative literature reviews, but not narrative reviews, violate the methodological precepts of behavior analysis by pooling data from nonidentical studies. We argue that it is impossible to assess generality without varying the context in which relationships are studied. Properly chosen data-aggregation strategies can reveal which behavior–environment relations are general and which are procedure dependent. Within behavior analysis, reluctance to conduct quantitative reviews may reflect unsupported assumptions about the consequences of aggregating data across studies. Whether specific data-aggregation techniques help or harm a research program is an empirical issue that cannot be resolved by unstructured discussion. Some examples of how aggregation has been used in identifying behavior–environment relations are examined.

In their critique of our quantitative review of studies of human choice (Kollins, Newland, & Critchfield, 1997), Baron and Derenne (2000) have made an impassioned case for caution regarding the pooling of data from different studies. Their present comments, like previous ones (Derenne & Baron, 1999), blend savvy articulation of methodological issues that apply to quantitative reviews of the literature with interpretations of the procedures and findings of our own review. These two agendas are tightly intertwined in both commentaries,¹ and for present purposes it is important to disentangle them.

With respect to our original review, it is gratifying to watch the self-correc-

tive processes of science in action. We regard the Kollins et al. (1997) review as a positive step toward integrating and understanding a scattered literature on human choice, but it would be astonishing if the review could not be improved upon. Baron and Derenne (2000) deserve credit for identifying some alternative approaches to conducting a review like ours, and we will welcome such efforts when they are forthcoming. The general conclusions of our review—that consequences matter in human choice, and that procedural variables modulate the matching relation between response allocation and relative reinforcer rate—are unlikely to change. It will be interesting, however, to determine the extent to which alternative ways of quantitatively summarizing the literature prompt conclusions that differ from ours about the modulators of human choice.

With respect to general processes of literature summary, we are pleased that Baron and Derenne (2000) now acknowledge that the benefits of the quantitative review “from the standpoint of precision and objectivity are

All authors contributed equally to the development of these comments.

Correspondence should be directed to Scott H. Kollins, Department of Psychology, Western Michigan University, Kalamazoo, Michigan 49008 (E-mail: scott.kollins@wmich.edu).

¹ Hereafter, for economy of expression, we use “Baron and Derenne” to refer collectively to the Derenne and Baron (1999) and Baron and Derenne (2000) commentaries. We specify year of publication when referring to specific articles.

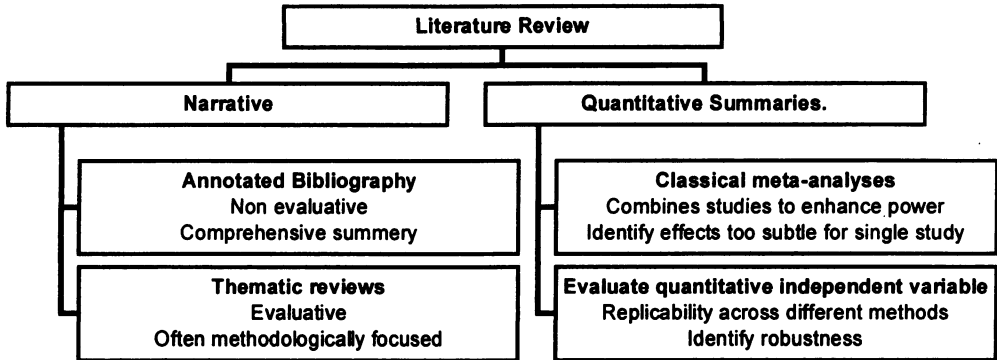


Figure 1. Strategies and tactics of literature review. Each of the two general strategies incorporates multiple classes of tactics, and each of these classes potentially incorporates a variety of specific techniques.

obvious" (p. 105), and now express interest in the refinement of this process for use with single-subject research. Nevertheless, we part company with Baron and Derenne on two important matters. First, although Baron and Derenne accurately identified some of the thorny technical and conceptual challenges that must be faced during the integration of data across single-subject studies, we believe that they erred in their pessimistic assessment of the implications of these challenges. Second, we believe that Baron and Derenne too narrowly characterized the options available to behavior analysts who seek generalities across related empirical studies. In the latter case, both broad strategies for seeking generalities across published studies (quantitative reviews and narrative reviews) can inspire a variety of specific tactics (as illustrated broadly by Figure 1). These tactics vary in their conceptual consistency with behavior-analytic methods and their heuristic value in behavior-analytic research. Baron and Derenne (2000), however, characterized the various quantitative review tactics as interchangeable, and on this point we disagree. Some methods of aggregating data are more consistent with behavior-analytic methods than others.

The present article amplifies our previous comments (Kollins, Newland, & Critchfield, 1999) on the importance of

(a) considering multiple approaches to the difficult task of integrating published studies and (b) discriminating among various quantitative review tactics, which may differ in terms of which data are combined, in what way, and to what end. For brevity and clarity of focus, we refer to our original review only when this informs a more general discussion, and instead address broader methodological issues raised by the Baron and Derenne commentaries.

NARRATE OR QUANTIFY?

The central goal of any literature review is to seek generalities across studies. A quantitative review employs some objective standard toward this end, and our original article evaluated a quantitative descriptor of behavior-environment relations (the slope of the generalized matching function) across contexts. To the extent that a quantitative descriptor remains stable across a wide range of conditions, one can be confident that it taps something fundamental. Alternatively, when the descriptor varies systematically across conditions, one may learn something about the controlling variables of behavior. Overall, the search for generality demands some variability across studies in subject characteristics (e.g., age, species), reinforcers, responses, settings, and so forth.

Baron and Derenne suggested limiting the raw material of quantitative reviews to studies with standardized methods in order to minimize the risk of confounding effects. For example, Baron and Derenne (2000) compared the reviewing process to an experimental investigation that would select "subjects [who] were similar in age and other individual characteristics, [and] that procedural details were the same except for the variable under investigation" (p. 105). But studies differ along many dimensions (e.g., the specifications of different brands of apparatus, the species or strain of non-human subjects, various methods of arranging control procedures, the exact values of independent variables, the sequence of conditions experienced by different subjects, the preexperimental histories of subjects, etc.). No useful collection of studies, in any topical literature, entirely avoids these procedural variances. A series of closely related studies from a single laboratory might meet the standards proposed by Baron and Derenne, but these studies would provide weak evidence of generality of effects. How, then, can literature reviewers reconcile concerns about confounding effects with the need to assess generality?

One solution is to accept the classic compromise, so often reflected in experimental research, between normative standards (methodological ideals) and the normal practice of science. From this perspective, research "standards" are viewed as guidelines to be approximated while each study is crafted to fit a unique set of situational demands and constraints; thus, research is viewed as a thoughtful problem-solving enterprise rather than a scripted ritual (e.g., Perone, 1999). A major responsibility of individuals who engage in this process is, of course, to acknowledge its limitations. Certainly, authors of quantitative reviews should take care to explain their methods and to note, in narrative form, likely sources of confounding as they perceive them. More formal tactics may be ap-

propriate as well. For example, we are intrigued by Baron and Derenne's (2000) suggestion that statistical methods of evaluating covariance might provide useful information about confounding effects.

An alternative approach is simply not to compromise. Perhaps the risks inherent in quantitative reviews outweigh the benefits, and narrative strategies should be employed exclusively. Viewing Baron and Derenne's proposal for conducting quantitative literature summaries as fixed standards, rather than guidelines, might prompt such a conclusion, because it seems unlikely that a useful quantitative review can be conducted according to these criteria. If confounding is to be avoided at all costs, then the review process is restricted to a compilation of nearly direct replications. More interesting compilations of studies would be viewed as unacceptable. Because Baron and Derenne (2000) devoted most of their article to illustrating the "incongruity in efforts to conduct quantitative summaries of single-subject studies" (p. 104), a reader could well conclude that, compared to the alternatives, quantitative literature reviews are especially riddled with pitfalls and thus to be avoided.

Such a conclusion would be unfortunate for several reasons. First, no blanket evaluation of quantitative methods is likely to do justice to the variety of circumstances under which these methods might be used. The utility of quantitative methods, like that of all methods, must be evaluated in context. For example, Baron and Derenne (2000) generally rejected between-groups methods, but conceded that "sometimes [between-groups comparisons] are demanded by the issue under investigation, as when age differences are the object of study" (p. 105). This is because, in selecting methods, an investigator must balance practical constraints inherent in the research question against the importance of answering that question. As Baron and Derenne noted, where experiments are

concerned, the ends help to dictate the means. Where literature reviews are concerned, the same considerations apply. The issue under investigation demands an answer, and it is an answer about phenomena in the aggregate, thus requiring compromises against ideal standards for inspecting the behavior of individuals under identical circumstances.

Second, it is axiomatic that a field should employ whatever methods are available to solve weighty problems such as summarizing findings across studies. Quantitative literature reviews are not deeply ingrained in the traditions of behavior analysis, so we argue in favor of broadening the field's armamentarium for summarizing its literature. Diversity in methods is important because no empirical strategy is foolproof, and therefore, science places a premium on converging evidence. The present exchange has focused on quantitative strategies, but it is widely recognized that narrative reviews, as well, are limited by their tendency to "rely heavily on the reviewer's subjective evaluations," which "differ markedly from review to review" (Baron & Derenne, p. 105). Fortunately, narrative and quantitative strategies of literature review are not mutually exclusive, and it makes sense to allow the two approaches to inform one another. For example, points of disconnect between our empirical findings (Kollins et al., 1997) and Baron and Derenne's narrative interpretations of the literature on human choice serve to focus attention on opportunities for research. We find it difficult to understand how this can be a bad thing, and predict that a combination of narrative and quantitative approaches will serve behavior analysts well in the integration of single-subject data.

Third, a false dichotomy looms large in Baron and Derenne's concerns about quantitative review strategies. Baron and Derenne worry that spurious findings can emerge when studies that are compared quantitatively along one dimension also differ on unexamined di-

mensions (see also Kollins et al., 1997). Quantitative reviews are not unique in this regard, however. To the extent that narrative reviews compare and contrast studies, they also incorporate subjective decisions about which procedural features of the studies to regard as trivial and profound. There is nothing controversial, therefore, in Baron and Derenne's (2000) assertion that "comparisons should be arranged in ways that minimize the confounding of variables" (p. 105)—and also nothing that, in principle, supports a preference for narrative over quantitative literature reviews.

AGGREGATION IS INEVITABLE

At the heart of our ongoing debate with Baron and Derenne appears to be disagreement over whether the data-aggregation practices of quantitative literature reviews really subvert essential conceptual and methodological traditions in behavior analysis. The quantitative review of Kollins et al. (1997) employed a technique called the empirical quantile-quantile (or percentile-percentile) plot to draw comparisons among groups of studies of human choice that differed on selected procedural dimensions. We argued that this technique preserves information about individual variation in behavior (unlike Fisherian inferential statistics) and emphasizes selected procedural differences across studies in order to identify potent effects (unlike conventional meta-analysis)—goals that are broadly consistent with the methodological foundations of behavior analysis (Kollins et al., 1999). Baron and Derenne (2000) responded that, in essence, data aggregation is data aggregation. Taken together, the two Baron and Derenne commentaries portray the data aggregation involved in all quantitative reviews as the brink of a slippery slope to poorer understanding of individual behavior.

Part of the appeal in Baron and Derenne's conservative position on quan-

titative literature reviews lies in its correspondence to an intuitive sense, shared by many behavior analysts, that data aggregation is a bad thing. Intuitions can be fallible, however, and in this case fail to recognize the fact that all scientists (even behavior analysts) engage in data aggregation at every step of the research process. At a broad level, the data-aggregation techniques of some kinds of quantitative reviews share properties with practices that already are common in behavior analysis. Consider that a response, as traditionally measured in the laboratory, incorporates numerous movements that may or may not share common causation, to say nothing of the fact that a single response represents the aggregate activity of billions of neurons and synapses. A response rate is a formal aggregate of responses, not all of which may share the same topography (e.g., Schick, 1971) or determinants (Schwartz & Gamzu, 1977). A condition mean aggregates events taking place over several sessions, ignoring a host of variables (motivational state, room temperature, subject handling, etc.) that can vary across days. Judgments of independent-variable effects for each subject require a joint consideration of events that occur in different conditions. And when scientists search for general effects—that is, patterns across subjects or groups within a study—they necessarily pool data from individuals who have not been treated identically.

Data aggregation across studies is unremarkable, except in the details of its sources of variance (e.g., programmed differences in experimental procedures). All data-aggregation techniques obscure variance and its sources, and do so by design. This is necessary because standardization is incomplete at all stages of the research process. Data-aggregation techniques aid in the search for relations that are robust enough to be detected above the noise that arises as a result. Without question, pooled data sometimes obscure important information about be-

havior (e.g., Baer, 1977; Johnston & Pennypacker, 1993; Michael, 1974; Sidman, 1960). But the fact that behavior analysts routinely (and comfortably) aggregate their data in selected ways is tacit acknowledgment that not all data-aggregation practices are created equal. At issue, therefore, is the basis on which data-aggregation techniques are to be included in, and excluded from, the methods of the field.

AVOIDING AGGREPHOBIA

Although behavior analysts long have predicted dire consequences of data aggregation (Baer, 1977; Johnston & Pennypacker, 1993; Michael, 1974; Sidman, 1960; see also the special section of *The Behavior Analyst* on statistical inference: Baron, 1999), their arguments, like those of Baron and Derenne, often have taken hypothetical or abstract form. In essence, these arguments help to establish that the sky might someday fall because of data aggregation, but they have little to say about whether it is currently falling. Behavior analysis is long overdue for a systematic examination of what might befall the field if data are aggregated in various ways.

What risks and benefits accrue from the data-aggregation practices associated with quantitative literature reviews? One source of answers may lie in the application of various quantitative tactics to the same data set to determine whether different techniques lead to discrepant impressions of the results. In this regard, Baron and Derenne's narrative comments about the Kollins et al. (1997) review are useful, but cannot substitute for new data or formal analyses that might substantiate concerns about our results. There is value in observing that procedures may have "opened the door to confounding effects" (Baron & Derenne, 2000, p. 103), because science proceeds by recognizing and ruling out alternative explanations. But this is not the same as demonstrating that any specific finding was illusory, or that any quantitative

technique is inherently flawed. Light can be shed on the human choice literature by replicating our quantitative integration using a different set of procedures for selecting studies, grouping studies, summarizing pooled data, and so forth. More important, light can be shed on the *general processes* of quantitative review through a comparative study of several quantitative procedures. This kind of study has the potential to illuminate the scope, magnitude, and ramifications of a variety of procedural dependencies in quantitative reviews. To our knowledge, no such investigation has been undertaken to date in behavior analysis.

Another source of insight regarding data-aggregation practices can be found in the form of research case histories. Here it is useful to ask which research programs have been distinctly altered by an infusion of data-aggregation practices, and in what ways. We are aware of cases (see Table 1) that might shed light on some of these questions. The table illustrates that data aggregation has produced both beneficial and unfortunate outcomes. We are not aware of any published evaluation of the issues involved that is systematic, comprehensive and, most important, tied closely to empirical evidence. Perhaps the present discussion will help to prompt such an analysis.

A particular advantage of case study methodology is that it may shed light on relevant human tendencies (e.g., judgment; see Perone, 1999) that the output of formal empirical evaluations of quantitative techniques cannot predict. For example, do data-aggregation practices exert different effects at different points in the developmental history of a research program? Do these effects depend on the style and sophistication with which investigators conceptualize their research problems? Do different techniques have different implications for socially driven processes such as theory construction and public policy formulation?

Overall, much remains to be determined about the implications of data

aggregation for behavioral science as practiced by behavior analysts. Presumably, the way in which aggregation takes place is of fundamental importance, rendering the answer to any question about the usefulness of data aggregation a form of *it depends*. The logical follow-up question—*depends upon what?*—demands a more systematic treatment than it has received to date in behavior analysis.

CONCLUSIONS

Baron and Derenne argued that no reasonable scientist should dismiss the pitfalls of pooling data. Neither, we submit, should any reasonable scientist ignore viable opportunities for seeking generalities across observations. Answers are so hard to come by in science, and generalities are so important, that investigators should seek insights wherever they are available. Literature reviews, both narrative and quantitative, can promote these insights. Quantitative review techniques, like all empirical endeavors, have their pitfalls. But good scientists know that methods dictate results; that no single finding is authoritative; and that generalities derived from any empirical exercise are worthy of recognition only insofar as they reflect converging evidence. In our view, these sensibilities minimize the liabilities associated with a quantitative literature review, while the potential benefits are left intact.

If nothing else, the present exchange serves as a reminder that current circumstances leave room for disagreement because the quantitative literature review is a historical rarity in behavior analysis and its parameters and implications remain to be explored. Thus, of far greater importance than the specifics of our original review is a reasoned approach to determining the best ways to evaluate general effects across studies, on any topic. Analyses like those suggested here could help behavior analysts to distinguish objectively among helpful, harmful, and innocuous data-aggregation practices, and thus to se-

TABLE 1

Some case studies of data aggregation

Consequences of data aggregation	Content area	Description	Representative papers
Data aggregation proved informative	Rate-dependent effect of drugs	Many studies indicated that the effects of certain drugs depended on the schedule that maintained behavior. Viewing drug effects as a function of response rates under control conditions revealed that rate, not reinforcement schedule, was the key factor.	Branch (1984), Dews and Wenger (1977)
	Matching law	A power function relates the ratio of behavioral allocation between two alternatives to reinforcement ratios across a range of species and situations.	Baum (1974), Davison and McCarthy (1988), de Villiers (1977)
Choice of aggregation tactics heavily influenced the course of a research area	Behavior under concurrent schedules	Herrnstein's (1970) matching law emphasized data aggregated over many sessions. Its popularity spawned a choice literature dominated by analyses of aggregate data from steady-state conditions. As a result, less is known about choice behavior in transition, the role of changeover contingencies, and events occurring in a single visit to a reinforcement alternative, factors emphasized or anticipated by Findley's largely ignored analyses.	Findley (1958), Houston and McNamara (1981), MacDonall (1998), Ziriak, Snyder, Newland, and Weiss (1993)
	FI scallop	Session averages produced the classic FI scallop, whereas within-session analyses yielded a different picture of behavior under this schedule.	Baron and Leinenweber (1994), Branch and Gollub (1974)
Data aggregation led researchers astray	Behavioral effects of food colors	Analyses of variance based on group means suggested that additives were behaviorally inert. Analyses of individual children identified stunning sensitivity that got lost in the grand mean.	Connors, Goyette, Southwick, Lees, and Andrulonis (1976), Weiss (1982, 1984, 1994), Weiss et al. (1980)
	Shock-maintained behavior	Session averages suggested that certain operants were maintained by response-contingent shock in some species. Molecular analyses by Galbicka and Platt revealed that the shock selectively punished long interresponse times.	Galbicka and Platt (1984), Morse and Kelleher (1977)

lect literature-review strategies based on informed cost-benefit analyses.

We have assumed, in our previous comments, that the benefits of employing objective (quantitative) standards toward summarizing single-subject literature are apparent to all. Baron and Derenne, like others before them (e.g.,

Salzberg, Strain, & Baer, 1987), have assumed that prohibitive costs are apparent as well. But the benefits and costs should be evaluated systematically rather than assumed. Speculation (on both sides of the argument) is cheap, and behavior analysis can do better. The matter cries out for the rigorous

empirical skills that behavior analysts have applied so successfully to other problems. There is, indeed, risk in exploring new methods. There is also risk in adhering too rigidly to traditional methods. The riskiest strategy of all, however, is to let subjective factors drive decisions about what methods best suit the analysis of behavior.

REFERENCES

- Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis*, 10, 167-172.
- Baron, A. (1999). Statistical inference in behavior analysis: Friend or foe? *The Behavior Analyst*, 22, 83-85.
- Baron, A., & Derenne, A. (2000). Quantitative summaries of single-subject studies: What do group comparisons tell us about individual performances? *The Behavior Analyst*, 23, 101-106.
- Baron, A., & Leinenweber, A. (1994). Molecular and molar analyses of fixed-interval performance. *Journal of the Experimental Analysis of Behavior*, 61, 11-18.
- Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22, 231-242.
- Branch, M. N. (1984). Rate dependency, behavioral mechanisms, and behavioral pharmacology. *Journal of the Experimental Analysis of Behavior*, 42, 511-522.
- Branch, M. N., & Gollub, L. R. (1974). A detailed analysis of the effects of *d*-amphetamine on behavior under fixed-interval schedules. *Journal of the Experimental Analysis of Behavior*, 21, 519-541.
- Conners, C. K., Goyette, C. H., Southwick, D. A., Lees, J. M., & Andrulonis, P. A. (1976). Food additives and hyperkinesia: A controlled double-blind experiment. *Pediatrics*, 58(2), 154-166.
- Davison, M., & McCarthy, D. (1988). *The matching law: A research review*. Hillsdale, NJ: Erlbaum.
- 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.
- deVilliers, P. (1977). Choice in concurrent schedules and a quantitative formulation of the law of effect. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior* (pp. 233-287). Englewood Cliffs, NJ: Prentice Hall.
- Dews, P. B., & Wenger, G. R. (1977). Rate dependency and the behavioral effects of amphetamine. In T. Thompson & P. B. Dews (Eds.), *Advances in behavioral pharmacology* (Vol. 1, pp. 167-228). New York: Academic Press.
- Findley, J. D. (1958). Preference and switching under concurrent scheduling. *Journal of the Experimental Analysis of Behavior*, 1, 123-144.
- Galbicka, G., & Platt, J. R. (1984). Interresponse-time punishment: A basis for shock-maintained behavior. *Journal of the Experimental Analysis of Behavior*, 41, 291-308.
- Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior*, 13, 243-266.
- Houston, A. L., & McNamara, J. (1981). How to maximize reward rate on two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior*, 35, 367-396.
- Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research*. Hillsdale, NJ: Erlbaum.
- 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.
- 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.
- MacDonall, J. S. (1998). Run length, visit duration, and reinforcers per visit in concurrent-schedule performance. *Journal of the Experimental Analysis of Behavior*, 71, 57-74.
- Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647-653.
- Morse, W. H., & Kelleher, R. T. (1977). Determinants of reinforcement and punishment. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior* (pp. 174-200). Englewood Cliffs, NJ: Prentice Hall.
- Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, 22, 109-116.
- Salzburg, 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.
- Schick, K. (1971). Operants. *Journal of the Experimental Analysis of Behavior*, 15, 413-423.
- Schwartz, B., & Gamzu, E. (1977). Pavlovian control of operant behavior. In W. K. Honig & J. E. R. Staddon (Eds.), *Handbook of operant behavior* (pp. 53-97). Englewood Cliffs, NJ: Prentice Hall.
- Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.
- Weiss, B. (1982). Food additives and environmental chemicals as sources of childhood behavior disorders. *Journal of the American Academy of Child Psychiatry*, 21, 144-152.
- Weiss, B. (1984). Food additive safety evaluation.

- tion: The link to behavioral disorders in children. In *Advances in clinical child psychology* (pp. 221–251). New York: Plenum.
- Weiss, B. (1994). Low-level chemical sensitivity: A perspective from behavioral toxicology. *Toxicology & Industrial Health, 10*, 605–617.
- Weiss, B., Williams, J. H., Margen, S., Abrams, B., Caan, B., Citron, L. J., Cox, C., & McKibben, J. (1980). Behavioral responses to artificial food colors. *Science, 207*, 1487–1489.
- Ziriax, J. M., Snyder, J. R., Newland, M. C., & Weiss, B. (1993). Amphetamine modifies the microstructure of concurrent behavior. *Experimental and Clinical Psychopharmacology, 1*, 121–132.