the same item for healthy females. Hence, the mean score for a healthy adult, sex unspecified, provides current data on the level of independence respondents considered "normal" at the time of the study.

Although the research presented by Festinger and Bounds is interesting, their critique of our study is less substantial and does not reflect a thorough analysis of the data. They inaccurately quote us as concluding that "social work students at the University of Minnesota do not have pronounced stereotypes about sex roles." [2] On the contrary, we said the following:

> Although the results of this research do not show the existence of a double standard when comparing healthy man and woman with healthy person, male and female subjects held a different concept of mental health for men and for women when the two sexes were compared directly. Thus, when the data were analyzed in this way, a double standard for mental health was substantiated. In addition, male and female subjects had a strong difference of opinion in their concepts of a healthy woman. These findings—which suggest that males and females have different expectations for a healthy, socially competent, adult woman—have important implications for potential difficulties in relationships between males and females at both personal and professional levels.[3]

In further criticizing our research methodology, Festinger and Bounds state, "Thus to present means for all the items thrown together, as is done by Harris and Lucas in all their tables, simply obscures whatever differences may exist." [4] This statement is erroneous. Mean scores for each item were calculated for the total population, male and female respondents separately, and, finally, undergraduate and graduate respondents separately. Charts containing these mean scores per item were not included in the published version of our article for the sake of brevity, but they are available on request. The analysis of items by total population indicates that on all but one item ("very skilled in business") respondents rated a healthy woman and a healthy man the same as a healthy person ($p \leq .05$). As we indicated in the article, male and female respondents reached a consensus on every item regarding characteristics for a healthy man, but 25 percent of the scored items showed a highly significant difference between male and female ratings for a healthy woman.[5] Content of these items is indicated by asterisks in Table 1 of the article.

Finally, we have received a large amount of international correspondence regarding the article. It is heartening to know that so much interest exists regarding research on sex-role stereotyping. As with most research in the human services, it is important to remember that results usually represent possible trends rather than proved facts and are most appropriately used to stimulate further discussion and exploration.

LINDA HALL HARRIS
Hopkins Public Schools,
Minneapolis, Minnesota

[2] Trudy Bradley Festinger and Rebecca L. Bounds, "Sex-Role Stereotyping: A Research Note," *Social Work,* 22 (July 1977), p. 314.
[3] Linda Hall Harris and Margaret Exner Lucas, "Sex-Role Stereotyping," *Social Work,* 21 (September 1976), p. 393.
[4] Festinger and Bounds, op. cit., p. 314.
[5] Harris and Lucas, op. cit., p. 392.

Statistical Evaluation of Clinical Effectiveness

In their article "Evaluating One's Own Effectiveness" (March 1977 issue), Bloom and Block argue that a therapist can and should determine whether or not some clinical intervention has been effective in terms of changing problem events in a desired direction. Specifically, a change resulting from intervention must be distinguished from a change that arises merely as a result of chance factors. As a tool for making such a distinction, the authors offer a statistical technique that involves comparison of an event's relative frequency following some intervention with the corresponding relative frequency of the event during a baseline period.

The goal of evaluating the effectiveness of a clinical evaluation is a laudable one. However, two problems exist that necessitate considerable questioning of the authors' conclusions—particularly of the exact probabilities that are claimed to emerge from their technique. Following a short description of our understanding of the technique, we will discuss these problems in turn.

To derive their method of statistical evaluation, Bloom and Block make use of the well-understood binomial distribution concept. A binomial distribution derives from a situation in which there are some number (N) of trials, each trial having some probability (p) of eventuating in a "success." Referring to the pedophilic fantasy example described by the authors, a "trial" consists of one intervention day, and a "success" is the occurrence of a fantasy frequency level below some predefined criterion. Therefore, $N=28$ (since there are 28 postintervention days). During the baseline days, two instances of pedophilic fantasy were low enough to be in the desirable range. Therefore, success probability is calculated to be, $p = 2/10 = .20$. The authors then characterize a "null hypothesis," which is that the intervention has no effect on the event. Stated statistically, the null hypothesis is that p, the success probability, remains at .20 during the intervention period, and any deviation of the observed proportion of successes from .20 is the result of random fluctuation.

Pitted against the null hypothesis is an "alternative hypothesis," which states that the intervention does have an effect, and that p thereby rises above .20 during the intervention period. To decide between the two hypotheses, the authors establish criterion-success frequencies that must be achieved for the practitioner to be able to reject the null hypothesis. In other words, if enough instances of the event occur at a desirable level during the postintervention period compared to baseline, then the clinician can feel reasonably confident that the intervention had a desirable effect.

The degree of confidence is stated at the .001 level. This means that, by

using this technique to determine whether or not one should reject the null hypothesis, the clinician has only a 1 in 1,000 chance of erroneously rejecting a null hypothesis that is actually true. In short, if the criteria are met, the authors suggest that the null hypothesis may safely be rejected in favor of the alternative hypothesis—and the intervention is thus deemed successful.

TABLE 1. THE NUMBER OF OBSERVATIONS OF A SPECIFIED TYPE (SUCH AS A DESIRED BEHAVIOR) THAT ARE NECESSARY DURING THE INTERVENTION PERIOD TO REPRESENT A SIGNIFICANT INCREASE (AT THE .001 LEVEL) OVER THE PROPORTION DURING THE PREINTERVENTION, OR BASELINE PERIOD

Proportions at Baseline	Number of Observations				
	12	16	20	40	68
.25	10 (9)[a]	13 (11)	16 (12)	28 (20)	48 (30)
.35	12 (10)	16 (13)	19 (15)	35 (25)	58 (37)
.45	— (12)	— (14)	— (17)	40 (29)	66 (44)

[a] Figures in parentheses are those included in Bloom and Block's original table.

PROBLEMS OF ANALYSIS

The first problem in Bloom and Block's analysis is that their use of the binomial distribution requires that all trials during the postintervention period be *independent* of one another—that is, that the probability of a success on one day in no way influences the probability of a success on any other day. Stated another way, data are not independent (that is, they *are* autocorrelated) if the ability to predict the level of an instance at a given time is enhanced by knowing what happened with past instances.[1]

Unfortunately, however, such is almost certainly not the case in virtually all clinical situations to which this analysis might be applied. Events on any given day are invariably related to events on nearby days. Therefore, the binomial techniques cannot appropriately be used. The foregoing problem is sufficient to cast serious doubt on the criteria that the authors present for a .001 significance level. To discuss the next problem, however, let us assume for the moment that independence is not an issue.

The second problem involves the probability that is established during baseline. The group of instances taken during the baseline period is, in fact, only a sample of the total population of instances that would be derived if the baseline measurement period were to be substantially (or indefinitely) long. Therefore, we need to distinguish between two quite separate probabilities.

First, we will define p to be the *actual* probability of some event during the baseline period—that is, the relative frequency with which the behavior has occurred over some long period of time. Next, we define p_B as the event probability that has been calculated using the sample of instances during baseline. Bloom and Block's reasoning rests on the assumption that $p=p_B$. But this is not necessarily true—and in fact will almost invariably be untrue. Rather, p_B is only an *estimate* of p. This fact must (and fortunately can) be taken into account when establishing the various criterion frequencies. Without going into the mathematics of the situation, a few remarks are in order.

First, since p is not known, but only estimated, there is uncertainty added to the situation over and above what is currently assumed by the authors. This means that the criteria necessary for a .001 (or any) significance level must be stricter than those presented by the authors in their Table 1. (P. 134.) An abbreviated version of that table is presented here (our Table 1), showing Bloom and Block's criteria along with the *actual* criteria—those that emerge when the uncertainty in baseline probability is taken into account.

Second, the greater the number of baseline instances on which p_B (the estimate of p) is based, the more accurate the estimate should be. (With an infinite number of such trials, p_B must equal p.) Since the accuracy of estimation depends on the number of baseline instances, there must accordingly be separate tables for various values of number of baseline instances as well as for various values of number of postintervention days. (Our Table 1 is for a baseline period of ten instances.)

A final issue is that the baseline instances represent only a sample of the population. Sometimes this sample will not produce even one occurrence of a rare problem event. When this happens, the authors implicitly acknowledge that the data are not representative and (in the context of their physical contact example) suggest that "because it was a part of the total problem, the worker was justified in assuming one instance out of the ten days of baseline." (P. 134.) However, it should be noted that if the observed baseline probability is zero then the assumption that it is equal to some specific nonzero value (.10 in the physical contact example) is totally gratuitous.

One could, with equal justification, assume that there were two instances during baseline, in which case the probability would have been .20, and the criteria would have been altogether different. Or suppose that there had been only 5 baseline days instead of 10. In such a situation, the one assumed behavior occurrence would similarly have dictated a probability of $1/5=.20$. Likewise, a baseline of 20 days would have produced a probability of $1/20=.05$, and so on. It is certainly the case that a zero-baseline probability raises problems for the authors' technique. But making up a probability out of thin air is worse than nothing—it gives the therapist totally groundless premises on which to make his or her evaluation decision.

[1] For a discussion of the transformation of autocorrelated data, *see* John M. Gottman and Sandra R. Leiblum, *How to Do Psychotherapy and How to Evaluate it* (New York: Holt, Rinehart & Winston, 1974), pp. 144–150.

The authors are dealing with difficult statistical issues, and as noted above, their goals are important ones. But the goals will not be optimally achieved by choosing to ignore the statistical difficulties and offering exact numbers as if such numbers derived from a problem-free technique. The two problems discussed above do not render Bloom and Block's method completely useless. The second problem—that p_B is only an estimate of p—is (with the exception of the zero-baseline issue) easily solvable, as exemplified in the present reply. The first problem—the independence issue—is not as easily gotten rid of, although methods are available for many situations.[2] In any case, therapists who use this technique should be cautioned that the numbers are not absolute—they may be useful as "ballpark figures," but their exact values cannot be taken very seriously.

GEOFFREY R. LOFTUS AND
RONA L. LEVY
School of Social Work
University of Washington, Seattle

[2] *See* ibid.

Bloom and Block Reply

We appreciate Loftus and Levy's response for its constructive advancement of our common cause in seeking better evaluation procedures for social workers in field settings.[1] Although there are some inaccuracies in their presentation of our procedure—for example, our method calls for comparison of the *number* of problematic events during intervention with the *proportion* of those events during baseline—we will focus on their central points and add some further thoughts on the practicalities of getting social workers to use evaluation procedures.

The first criticism involves the assumption of independence between events being measured. This is an excellent point, previously mentioned in the original source of the procedure.[2] However, *not* using a statistical procedure because it includes statistical assumptions about which reasonable people may disagree has, perhaps, greater negative consequences than the risks involved in using that procedure. Foremost among the consequences is that unless some easy-to-use evaluation methods are available, individual social workers will most likely not evaluate their practice, leaving the evaluation field to the more complex, expensive—and rarer—control group studies with all of the assumptions they make in translating laboratory designs into the community.

We agree with Hersen and Barlow's observation that repeated individual measures help to search for sources of individual variability, and thus make a contribution to the question of what is related to what in the individual case.[3] This is the real point behind the assumption of independence of events, and by taking the risk, we may be advancing our knowledge. However, we agree with Loftus and Levy's implicit point, that everyone should be cautious in using any statistical procedures; the user should ask for appropriate critical information about the method.[4] But most important, everyone should be willing to take the necessary and *inevitable* risks involved in evaluating live cases in the field setting, lest this become a dead profession.

The second major point concerns estimating baseline conditions. This is a complex issue about which we have tentatively come to some working decisions in order to get on with the business of evaluation without the presence of computers or teams of researchers. The problem begins with the task of selecting events to be measured. A critical, though rare, event (such as the pedophilia discussed in our article) in effect demands to be measured; we must evolve some measurement procedure in order to be informed about the efficacy of our intervention.

Questions such as how stable a picture must be obtained in baseline before intervention is to be started may be taken out of our hands, although we would ideally like extensive, stable measures of the event before we start treatment. By making some objective and repeatable measurement rules the basis for our inferences of practice efficacy, we can begin the evaluation process. Researchers may disagree with the assumptions and rules we have made, but at least they are reasonably clear how we made them and why. Unfortunately, we are not clear about the alternative figures Loftus and Levy propose in their Table 1. We presume, however, that if more space were available they could present their rules to would-be users for consideration.

Which set of rules among the many possible is to be preferred should itself be the subject of empirical study. We are currently engaged in several studies that seek to expand the usefulness of our evaluation procedures in special situations, including those we label $N = 1$, $T = 1$ (single-subject design with onetime contact); $N = 1$ and several (making simultaneous comparisons between and among individuals, pairs, and groups in group contexts); and $N = 1$, $T = \infty$ (long-term single-subject design suitable for use in preventive studies). The point is that we are not satisfied with the evaluation procedure as it is, and we are trying to work out some of the problems. We hope further constructive interchanges among researchers and practitioners will clarify this procedure and ultimately lead to better solutions.

MARTIN BLOOM
School of Social Work, Washington University, St. Louis, Missouri
STEPHEN R. BLOCK
Department of Social Service, Indiana University Medical Center, Indianapolis

[1] *See* Rona L. Levy, "Single-Subject Experimental Designs in Social Work Education." Paper presented at the Council on Social Work Education Annual Program Meeting, Phoenix, Arizona, 1977.

[2] Martin Bloom, *Paradox of Helping: Introduction to the Philosophy of Scientific Practice* (New York: John Wiley & Sons, 1975), p. 201.

[3] Michael Hersen and David H. Barlow, *Single Case Experimental Designs: Strategies for Studying Behavior Change* (New York: Pergamon Press, 1976).

[4] Bloom, op. cit., p. 198.