Psychological Methods
1999, Vol. 4, No. 3, 243-249

1082-989X/99/$3.00

How to Show That 9 > 221: Collect Judgments in a Between-Subjects Design

Michael H. Birnbaum
California State University, Fullerton

In between-subjects (BS) designs, different groups may be asked to make judgments on numerical rating scales. According to judgment theory, judgments obtained BS are not an ordinal scale of subjective value. This article illustrates how BS designs can lead to strange conclusions: When different groups judge the subjective size of numbers, 9 is judged significantly larger than 221. The theory is that 9 brings to mind a context of small numbers, among which 9 seems "average" or even "large"; however, 221 invokes a context of 3-digit numbers, among which 221 seems relatively "small." Within-subjects, however, judges would not have said 9 > 221. Implications of this problem and suggestions for dealing with it are discussed.

The purpose of this article is to illustrate how between-subjects (BS) experiments, in which the dependent variable is a judgment, can lead to dubious conclusions. Although this point has been made previously (Birnbaum, 1974, 1982, 1992; Birnbaum & Mellers, 1983; Greenwald, 1976; Grice, 1966), the implications of this thesis may not yet be fully appreciated by researchers. This article uses a simple example to illustrate how difficult it is to compare judgments between subjects.

When different groups of people judge a stimulus, the response by a given person on a specific occasion is theorized to be a function of subjective value:

$$R(i,k) = J_k(s_i) \quad (1)$$

Where $R(i,k)$ is the response to stimulus i in context k; J_k is the monotonic judgment function that maps subjective value into responses in context k for that person; and s_i is the momentary subjective value of stimulus i on that occasion. When there are different people in different contexts, one can also add subscripts for individuals to all of the variables.

This research was supported by Grant SBR-9410572 from the National Science Foundation.

Correspondence concerning this article should be addressed to Michael H. Birnbaum, Department of Psychology, California State University, P.O. Box 6846, Fullerton, California 92834-6846. Electronic mail may be sent to mbirnbaum@fullerton.edu. More information about this article can be found on the World Wide Web at http://psych.fullerton.edu/mbirnbaum/home.htm.

Even though the judgment function might be monotonic within each person, Equation 1 does not guarantee that BS judgments are an ordinal scale of subjective magnitude; that is, $J_1(A) > J_2(B)$ does not guarantee that $A > B$. It can easily occur that $A < B$ but $J_1(A) > J_2(B)$. In other words, the subjective value of A is less than that of B, but B can receive a higher rating in its context than A does in its context.

Research in which context is systematically manipulated between groups has shown that the same stimulus may be judged "large" or "small," depending on the other stimuli that form the context for judgment (Parducci, 1965, 1968). Excellent summaries of hundreds of such experiments are presented by Parducci (1995) and Poulton (1989). Parducci's (1965) range-frequency theory describes contextual effects in judgment in terms of the distribution of stimuli that form the context for judgment. The range and frequency principles are illustrated in Figures 1 and 2. Actual judgments are represented as a compromise between these two principles (Parducci, 1965, 1995).

Because the same stimulus can get different judgments by different groups, a BS comparison of judgments does not allow us to compare the stimuli. For example, Birnbaum (1974) reported that the number 450 received a higher judgment in a positively skewed distribution of numbers than the number 550 received in a negatively skewed distribution of numbers. It would be wrong to conclude from this finding that people think 450 is greater than 550, because within subjects (WS), no individual gave a higher judgment

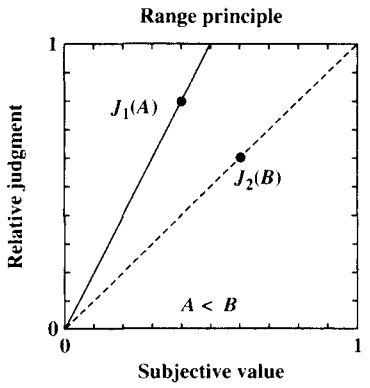

Figure 1. Illustration of the range principle. In each context, the minimum stimulus receives the minimum response, and the maximal stimulus receives the maximal response. Stimulus A is lower in subjective value than B; however, Stimulus B is presented in a context in which the maximum stimulus is higher, so A is judged (between subjects) higher than B. For example, if Stimulus A is the number 9, it could receive a higher judgment in its context than 221 if the context of 9 has a smaller maximum than the context of 221.

to 450 than to 550. The conclusion holds only BS (see Birnbaum, 1974, Figure 2).

According to Birnbaum's (1974, 1982) treatment of Parducci's (1965, 1995) range-frequency theory, in BS designs, different people who experience different contexts have systematically different J functions in Equation 1. Although each J function might be strictly monotonic WS, averaged BS judgments are not even a function of subjective value, and therefore, BS mean judgments are not even an ordinal scale of subjective value.

The direction of empirical effects can be opposite in WS and BS designs (Birnbaum, 1975; Grice, 1966). For example, Jones and Aronson (1973) studied judged fault attributed to rape victims as a function of their respectability. Respectability was manipulated between groups. A victim described as a "virgin" or "married" before the rape was rated on average more at fault than a victim described as a "divorcee." Jones and Aronson (1973) theorized that this result is due to belief in a "just world," in which people get what they deserve. A respectable person would not deserve to be a victim, so she must have done something wrong (she must be at fault) to deserve it. However, this result, which can be replicated BS (Birnbaum, 1982, Figure 17.23), is reversed WS. Birnbaum (1982, Figures 17.22–17.24) reported that in WS designs, the divorcee is judged most to blame. In other words, WS and BS research yield opposite conclusions on the judged fault of victims.

Birnbaum (1982, Figure 17.25) explained these conflicting results using an extension of range-frequency theory. The theory explains why BS and WS designs give different results. According to the range-frequency interpretation, all judgments are made relative to some context, including judgments of fault. When asked to judge the fault of a victim who was a virgin, the judge needs to know the context. In other words, "how much is she to blame—compared to what?" In a BS design, the context for those who judge the virgin victim may be other virgins. For the group who judges the fault of the divorcee victim, the context may be other divorcees. Thus, the divorced victim is less to blame relative to divorcees than the virgin victim is to blame in comparison with other virgins; however, the divorced victim is more to blame than is the virgin when they are compared with each other. Thus, BS ratings can be in the opposite direction from those in WS ratings, which are made relative to a context that includes both types of victims.

Figures 1 and 2 illustrate how the range and frequency principles predict when WS and BS judgments will give opposite conclusions. Note that in each figure, $A < B$, but comparing judgments between groups who received different contexts, the judgment of A exceeds the judgment of B. In this analysis, each

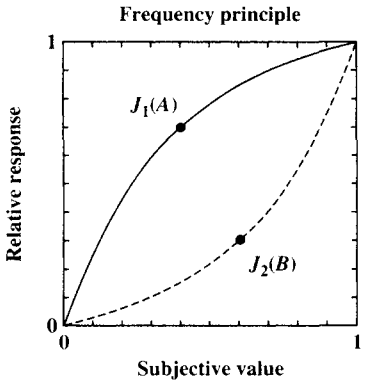

Figure 2. Illustration of the frequency principle. Stimulus A is subjectively lower than Stimulus B but can receive a higher judgment than B. A is presented in a positively skewed distribution, and B is presented in a negatively skewed distribution with the same end points. Because A has a higher cumulative probability in its context (in its distribution) than B has in its context, A is judged higher than B between subjects. For example, 9 could be judged higher in its context than 221 in its context if the end points were the same, but 221 invoked a context that included a greater relative frequency of larger numbers than did 9.

judgment function (each curve in the figures) is strictly monotonic (so WS, the virgin is less to blame than the divorcee). However, BS, the opposite conclusion will be reached (the virgin is more to blame than the divorcee).

Research on the so-called "base-rate fallacy" also leads to very different conclusions when studied BS and WS. In BS designs, base rate has small effects in Bayesian inference problems (Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1973), leading to the conclusion that people ignore or neglect the base rate. In WS designs, however, base rate has a large and systematic effect, contradicting the notion that people disregard base rate (Birnbaum & Mellers, 1983).

Unfortunately, research topics with content and history also contain many side issues that complicate the discussion of well-established principles of judgment. In social psychology, it is often argued that if we allow people to know what we are studying, they might behave differently because of "demand" characteristics of the situation. Therefore, it is argued, we should use BS designs in order to prevent participants from knowing what variable has been manipulated (Nihm, 1984).

In base-rate-fallacy research, some argue that the world "is more like" a BS than a WS design (Kahneman et al., 1982) and that the basis for generalization is similarity of an experiment to the domain of intended generalization. Brunswik (1956) argued that the principles of sampling of participants should also be applied to sampling of experimental situations, based on the argument that the basis for generalization is the representativeness of an experiment to the natural ecology. Others dispute this position, arguing that the basis for generalization is theory and that ideas of similarity (representativeness) and random sampling are the weakest forms of "theory" available. Although Brunswik argued against BS designs, it is ironic that notions developed from his ideas have been used to argue for the opposite conclusion. These philosophical and content-related arguments have confounded discussion of the core issues.

A third argument is also offered as follows: If we do nothing else between groups besides present different stimuli, we can assume that the context (and hence, J function) is the same between groups because of the random assignment of people to conditions (Poulton, 1989). The present study refutes that assumption.

The present study uses judgments of the size of numbers to illustrate that judgments in a BS design are not an ordinal scale of subjective magnitude. This study is intended to show that a clearly "wrong" conclusion is reached when we compare judgments obtained in a BS design. My purpose is to use an obvious example in a simple situation to encourage people to take the issue seriously when it arises in complex situations where the correct answer is not obvious. I conjectured that 9 might easily be judged to be a "bigger" number than 221 in a BS design. If the number 9 invokes a context of 1-digit numbers, then 9 would seem "larger" in its (hypothesized) context than would 221, if 221 invokes a context of 3-digit numbers.

Method

Over the Internet (URL, http://psych.fullerton.edu/mbirnbaum/exp.htm), people were recruited to serve in a "1-minute judgment study." The experiment was also advertised by E-mail sent to members of the Society for Judgment and Decision Making.

People were assigned to either the 9 or the 221 conditions by the following procedure. Those accepting the invitation to participate were asked to click on their birth month. Odd months (January, March, May, etc.) were assigned to the 9 condition, and even months (February, April, etc.) were assigned to the 221 condition. After 40 people participated, this association was reversed, to counterbalance any effects of birth month. By plan, the experiment was completed when there were at least 40 participants in each group; there were 45 and 40 in the 9 and 221 conditions, respectively.

The instructions asked participants to indicate their E-mail address, nation of birth, sex, education level, and name (optional) and to "judge, how large is the number 221?" or "judge, how large is the number 9?" Judgments were made on a 10-point scale, ranging from *very very small* to *very very large*. The form also requested comments (optional). The only difference between conditions was the number to be judged (9 or 221).

Results

The number 9 received a mean judgment of 5.13, and the number 221 received a mean judgment of 3.10. This difference is statistically significant, $t(83) = 3.52, p < .001$. Therefore, the BS conclusion is that the subjective size of the number 9 is greater than that of 221. Note that the mean judgment of 9 is near the

"average" of numerical size, whereas the number 221 is judged toward the "small" end of the scale. Responses ranged from 1 to 10 in each condition, and the standard deviations were 3.58 and 1.88 for 9 and 221, respectively.

The difference between 9 and 221 is so strong that it was statistically significant, even when data were analyzed with only the first 10 people in each condition. The same relation between means holds within each sex and within each education level. For the 43 women in the study, the mean judgments of 9 and 221 were 5.76 and 2.78, respectively. For the 42 men, means were 4.35 and 3.36 for 9 and 221, respectively. The same relation between means was also found for those with high school, college, master's, and doctorate degrees. Thus, the same BS conclusion holds when data are partitioned by sex and by education: The number 9 consistently receives a "larger" mean judgment than does the number 221.

Discussion

Because the number 9 is judged to be significantly "larger" than 221 in a BS design, but we don't really believe that people think $9 > 221$, we should be skeptical of any other BS conclusion based on the same type of study. Remember, no individual ever said 9 is greater than 221. The conclusion holds only between groups who were presented only one number for judgment.

These data show that assumption of random assignment to conditions does not guarantee that the J functions in the two groups are the same.

These results can be explained by the theory that different stimuli bring with them different contexts, that is, that context and stimulus are confounded in BS designs. Different contexts produce different judgment functions. If people in the 9 condition think of a smaller context than do the people in the 221 condition, then one can explain the results with the range and frequency principles, as illustrated in Figures 1 and 2. Figure 1 shows that $A < B$, but the judgment of A exceeds that of B, because the context in Group B has a greater subjective maximum than the context in Group A. Presumably, those in the 221 condition imagined a larger maximum stimulus than did those in the 9 condition. The number 9 might also invoke a context with a greater frequency of small numbers in its context, which would also predict the relation observed, as illustrated in Figure 2, even if the subjective end points were the same.

Data collected BS typically have higher variance than those collected WS (e.g., see Birnbaum & Mellers, 1983, Figure 2). The high variability of judgments in each BS group can be explained by the theory that different individuals within each group imagine different contexts when making their judgments. For example, if a person imagined all positive numbers as the context, then that person would judge 9 to be "very small." Another person who thinks of all numbers from negative infinity to positive infinity, would judge 9 to be "average" in size. When a person considers the context of numbers from 1 to 10, then 9 is judged to be "very large." It is theorized that those in the 9 condition were more likely to think of a context of small numbers than did those in the 221 condition, producing enough of a change in context to override the actual relation between the sizes of the numbers.

Some might argue that there was "no" context in our experiment, because only one stimulus was presented for judgment (Poulton, 1989). In contrast, it can be argued that when the context is not specified, it permits many contexts, confounding stimulus and context (Birnbaum, 1975, 1982, 1992; Greenwald, 1976). By presenting different stimuli to different groups, different contexts were created. Without theorizing that the contexts are different, it is hard to see why 9 would be judged significantly greater than 221; however, the theory that different contexts are invoked by different stimuli makes it easy to understand the results.

Would "Anchors" Solve the Problem Between Groups?

An "anchor" is a stimulus that receives the same judgment in both contexts. For example, we might instruct people to assign the category *very very small* to 1 and the category *very very large* to 1,000—would such a procedure guarantee the comparability of responses between groups? The answer is no; $J_1(x_1) = J_2(x_1)$ and $J_1(x_2) = J_2(x_2)$ does not guarantee that $A > B$ if and only if $J_1(A) > J_2(B)$.

If J functions were always linear (e.g., if judgments were determined only by the range principle), then two "anchors" would indeed anchor the scale. However, the frequency principle implies that the derivative of the J function is also proportional to the contextual stimulus density function (Birnbaum, 1974), producing nonlinear J functions that will not be anchored by "anchors."

Birnbaum's (1974, Figure 2) data illustrate that an-

choring the lowest and highest points of the scale does not guarantee that the *J* functions coincide. In that study, different groups of people judged the sizes of numbers that ranged from 108 to 992. All participants assigned their lowest response to 108 and their highest response to 992. However, between groups, 450 was judged significantly "larger" than 550. The frequency principle, illustrated in Figure 2, was used to design the experiment to create this effect. The frequency principle also explains the results in Birnbaum's (1974) Figure 3, which showed two contexts in which the judgment functions coincide at three points (both end points and the midpoint), and yet there are two regions in which smaller numbers are judged significantly higher than larger numbers. See also Parducci (1995).

Birnbaum's (1974) Figure 4 shows that even when the judgment functions coincide at 4 points, the *J* functions can be systematically different, and smaller numbers can still be judged significantly bigger than larger numbers. Thus, the use of anchors to supposedly "pin down" the *J* function at 2, 3, or 4 points does not guarantee that the *J* function is also the same at all points. The data of Birnbaum (1974) demonstrate that the failure of "anchors" to anchor the scale is not only a possibility in principle, but also that anchoring does not work in practice.

Implications for Applied Research

Although the effects of context have been tested, demonstrated, and replicated in many experiments with a wide variety of stimuli and judgment scales, applied research has not always heeded the lessons of experimental research. In the Soviet Union, between-group questionnaires sent to managers of farm collectives showed that now-discredited farm practices were improving grain yields at the same time that total grain production in the nation was decreasing (Medvedev, 1971). Unfortunately, important applied problems are still being studied by the use of judgments obtained BS. In current applications, for example, people are asked to rate how satisfied they are with a medical treatment, a health care system, or a teacher. Decision makers then compare judgments between groups who have not experienced the other treatment. The danger is that results may be opposite from those that would be obtained if each person compared treatments, health systems, or teachers.

Consider the method typically used to evaluate teaching performance of university professors. Judgments of teaching effectiveness are obtained from different groups of students who take different instructors in different classes. Different instructors use different materials, have different personalities, use different standards for grading, and differ in any number of additional contextual variables. Unlike the present study, in which people were randomly assigned to conditions, students are not randomly assigned to instructors or classes but select themselves into majors, courses, and instructors. Instructors are not randomly assigned to departments or courses.

Student evaluations are well-known to have high variability: A given teacher receives a wide range of ratings by different students, and the same teacher can receive very different mean ratings in successive semesters in the same class (Greenwald, 1997). Such high variability suggests that different confounded contexts are at work.

The problem is that the student's context for judgment is completely confounded. The student does not know how other instructors would teach the same class or if this class has presented the proper content of the course. The student does not know what grade he or she would have achieved on a nationally standardized test of content. The student's context may be affected by the student's own grade. Any interpretation of student evaluations is speculative, because no one has ever done the study properly. A proper study would include a complete, counterbalanced, WS design in which the same students would take all classes at a university from all instructors and rate them all afterward without knowing their grades.

In response to criticisms of methods of applied research, applied practitioners sometimes make the following argument. Applied work need not be based on theory backed by experimental evidence, and applied methods should be considered innocent until proven guilty in each applied context. I argue the opposite: Applied practices should be based on sound theory and shown to lead to correct conclusions before they are used to make important decisions.

To compare two professors, we could have the same students take both sections of the same class, with different students randomly counterbalanced in the order of taking the different professors, and they should make their judgments without knowledge of grades. That would be hard to do but not impossible. One could also use a panel of professors who would audit many classes and evaluate samples of teaching quality provided by different instructors. Although expensive, this procedure is also feasible. One can also compare instructors in the same course

by objective, standardized tests of student achievement.

For medical systems, proper research would allow the same people to experience both HMO and the fee-for-service systems before comparing them.

Double-Blind Treatment Experiment

The classic, double-blind, BS treatment study does not escape the criticism of this article if the dependent variable is a judgment. In this classic design, patients are randomly assigned to receive the treatment or placebo, and neither the doctor nor the patient knows which treatment has been administered. Although this elegant experimental design eliminates two problems of student ratings (confounded assignment and lack of blindness with respect to grades), it can lead to wrong conclusions if the dependent variable is a judgment.

Suppose the *experimenter* were asked to evaluate the effects of a treatment by studying only one group (either the treatment or control group). The logic of an experiment dictates that the experimenter must have data for two groups in order to estimate the treatment effect. The experimenter needs both groups and needs to know which group is which in order to estimate the effect of the treatment.

However, when we ask people who experience only one treatment (or placebo) to judge the effect of that treatment, we ask them to judge something that even the experimenter could not judge *in principle.* The experimenter cannot judge the effect of the treatment, even with data for an entire group and knowledge of which group it was. No participant in the study who experiences only one treatment can answer the question, just as the experimenter cannot answer the question without data from both groups, yet the study hopes to learn something new by combining judgments from many people, each of whom cannot, in principle, answer the question. Thus, even in the best type of double-blind experiment, if the dependent variable is a BS judgment, the comparison of judgments between groups is not valid.

This Criticism Does Not Apply With Objective Dependent Variables

My criticism does not apply to BS experiments with objective dependent measures. For example, the criticism would not apply to a BS educational experiment in which the dependent measure was a standard test of scholastic achievement. Nor would this criticism apply to a study of antibiotics with dependent measures such as days with fever or deaths.

A study of a cold remedy could also solve this problem by using judgments in a WS design, counterbalancing the order of treatment and placebo. For example, one could ask each person to try Pill A for one cold and Pill B for another. Each patient would judge which pill seemed to work better. People could be randomly assigned to factorially counterbalanced orders (instructed to use Pill A for the first cold and B for the second, or vice versa), with labels "A" and "B" counterbalanced. Both patients and doctors should remain blind to whether Pill A or Pill B contains the active ingredient. The dependent variable could be a comparative judgment of which cold was least severe. A WS design puts the participant in the position of an experimenter with two groups.

WS experiments have their own contextual effects, which are probably better understood than those in BS experiments. For example, there may be sequential effects or memory effects that influence the comparison of two colds; additionally, the labels "A" and "B" might bias the judgments. For this reason, one uses counterbalanced designs with theoretical models to disentangle the effects of the treatment from those due to such factors as the order of the treatments or labeling.

Another Solution: Use a Triple-Blind Study

The criticisms of BS judgments would also not apply to BS experiments in which independent judges rate the dependent measure for both groups (and also remain blind to the experimental treatment). This design may be called a triple-blind study. The participants (e.g., patients) are blind with respect to the treatment; the person who administers the independent variable (e.g., the doctor) remains blind; finally, the judge of the dependent variable remains blind with respect to the treatment and evaluates the dependent variable for both the treatment and control conditions. Although patients are assigned to treatments BS, judges evaluate the results WS.

For example, suppose one wanted to assess the effectiveness of a new ingredient for wrinkle cream. In order to assess the ingredient, it may be necessary to use judgments of the "luster" of faces of people in the study. The plan is to compare those who were randomly assigned to use the experimental cream (with the active ingredient) against those who received the placebo cream (without the active ingredient). Both patients and doctors should remain blind to the condition. The judges who evaluate the facial skin of patients should also be blind with respect to the treat-

ment. However, the *same judges should rate patients from both groups.*

If different groups of judges evaluated the faces of those who received the experimental cream and those who received the placebo, however, then the research might reach the wrong conclusions. According to range-frequency theory, if the cream were effective only for the worst (lowest luster) cases, and if it had little effect in other cases, then if the distribution of skin condition were normal in the placebo condition, the distribution would be positively skewed in the treatment condition. If judgments were made between groups, range-frequency theory predicts that the cream would incorrectly appear harmful (mean judgments would be lower for the treatment), even though it was actually beneficial. If the same judges rated both groups, however, range-frequency theory predicts that the treatment would be correctly determined to be effective (mean judgments would be higher).

Conclusions

If people had been asked to compare 9 and 221, they would have judged 9 < 221. If you agree that 9 is not greater than 221, you should be skeptical of studies that use methods that yield the silly conclusion that 9 is significantly "bigger" than 221. I chose this study of numbers because it does not tempt us to construct deception and "demand" explanations for the WS result, nor is it appealing to argue that the world is really "more like" a BS design when it comes to numbers. All of the participants have experienced numbers both larger and smaller than the numbers used here. The key to the result is that when judges are "free" to choose their own contexts, they choose different contexts for different stimuli. For this reason, it is important to beware of conclusions based on judgments obtained between groups of people who experienced different contexts. Even when there is "no" context besides the stimulus itself, comparison of judgments between subjects can be misleading.

References

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15,* 89–96.

Birnbaum, M. H. (1975). Expectancy and judgment. In F. Restle, R. Shiffrin, N. J. Castellan, H. Lindman, & D. Pisoni (Eds.), *Cognitive theory* (Vol. 1, pp. 107–118). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1992). Should contextual effects in human judgment be avoided? [Review of E. C. Poulton, *Bias in quantifying judgments*]. *Contemporary Psychology, 37,* 21–23.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45,* 792–804.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 83,* 314–320.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186.

Grice, G. R. (1966). Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin, 66,* 488–498.

Jones, C., & Aronson, E. (1973). Attribution of fault to a rape victim as a function of respectability of the victim. *Journal of Personality and Social Psychology, 26,* 415–419.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Medvedev, Z. A. (1971). *The rise and fall of T. D. Lysenko.* Garden City, NY: Anchor Books.

Nihm, Sue-Doe (1984). Self-reports on mental processes: A response to Birnbaum and Stegner. *Bulletin of the Psychonomic Society, 22,* 426–427.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Parducci, A. (1968). The relativism of absolute judgment. *Scientific American, 219,* 84–90.

Parducci, A. (1995). *Happiness, pleasure, and judgment.* Mahwah, NJ: Erlbaum.

Poulton, E. C. (1989). *Bias in quantifying judgments.* Hillsdale, NJ: Erlbaum.

Received July 24, 1998
Revision received April 14, 1999
Accepted April 21, 1999 ■