JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1983, 39, 499-501 NUMBER 3 (MAY)

COMMENT

MATCHING, STATISTICS, AND COMMON SENSE

Two papers have appeared recently, one by Mullins, Agunwamba, and Donohoe (1982) and one by Wearden and Burgess (1982), disagreeing with conclusions of my 1979 paper on choice experiments (Baum, 1979), the former on the basis of argument, the latter on the basis of additional data. Both these papers appear mistaken; the conclusions of my earlier paper still stand.

Mullins et al. criticize my discussion of variations in the exponent a of the power law of behavior allocation:

$$\frac{B_1}{B_2} = b\left(\frac{r_1}{r_2}\right)^a .$$

I considered separately sets of data in which B_1 and B_2 were measured by counting discrete responses and sets of data in which they were measured by time spent at the alternatives. I found both undermatching ($a < 1.0$) and overmatching ($a > 1.0$), as well as matching ($a = 1.0$). Undermatching was the more common deviation from matching, by far.

Mullins et al. make three claims: (1) that I incorrectly concluded that the time ratios generally conformed to the matching law, (2) that the methods I used to assess deviations from matching were faulty, and (3) that my suggesting a range of slopes that might be considered to approximate matching was mistaken. I consider all three claims incorrect. I will consider them in order.

MODES, SKEW, AND SCALE

The first claim disputes my description of the frequency distribution of power-law exponents (slopes in log-log coordinates) from ratios of times spent at two alternatives. Whereas the distribution of slopes from response ratios was roughly symmetrical with a mode, median, and mean clearly less than 1.0, I described the distribution of time-derived slopes as skewed toward values less than 1.0, but having a mode at about 1.0. The skew reflects a higher frequency of undermatching than overmatching among those data sets. The asymmetry leads also to a median and mean less than the mode at 1.0. For a skew-independent estimate of central tendency, the mode (provided a clear mode is present) is the best.

Attempting to argue that the central tendency fell short of 1.0, Mullins et al. make statements such as "two thirds of the observed slope values for the time measure are less than 1.0." They go on, "This departure from symmetry with only 17 values greater than 1.0 is statistically significant ($p < .02$)" (p. 324). Although no hint is given as to what statistical test was used or what assumptions it entailed, these statements indicate nothing more than that the distribution was skewed.

Can something be said about central tendency? Mullins et al. try, by redoing the frequency distribution using smaller class intervals than I used. Indeed, they find that a clear mode appears at about .9. But anyone who has ever struggled to find the best representation of a frequency distribution knows that playing with class intervals can be a dangerous game. Using smaller class intervals means fewer data per class interval, spreading out the data, and less reliable estimation of the mode. Indeed, if the class intervals are too small, one finds more than one mode. Although Mullins et al. make no mention of it, their class intervals were small enough to produce a multimodal distribution. Together with the mode they mention at .9, there are two lesser modes at .7 and 1.1.

Using large enough class intervals to produce a unimodal distribution results in a modal class interval that includes 1.0. I preferred the single mode, and still do, for two reasons: (1) I see no sense in more than one mode, and (2) none of the slopes between .9 and 1.1 was, by any obvious criterion, reliably different from 1.0.

Mullins et al. make another error in analyzing the variation in slope: Instead of representing the slopes on a logarithmic scale, as I did, they use an arithmetic scale. An arithmetic scale is appropriate to display and compare differences. Slopes and exponents, however, have the properties of ratios. Factors give the relationships among them, not differences. A slope of .4 differs from a slope of .8 as much as one of 1.6 differs from one of .8. A slope of 1.2 is closer to a slope of .8 than is a slope of .4. The basic reason for this is that the inverse of any slope or exponent is its reciprocal. Since 1.0 equals its own inverse, it represents the center of the scale, half of which lies between 0 and 1.0, and half of which lies between 1.0 and ∞. The logarithmic transformation has two desirable effects: It renders the scale symmetrical around 1.0, and it represents factors as distances, so that equally different slopes appear equally far apart. To get a correct idea of the relationships among different slopes, particularly degrees of undermatching and overmatching, one should use a logarithmic scale.

The paper by Wearden and Burgess (1982) illustrates how far one can be misled by the use of inappropriate scale. They gathered slopes from data sets published between 1977 and 1979, constructed frequency distributions, and also compared slopes for time ratios with slopes for response ratios in those experiments where both measures were taken. Like Mullins et al., they represented the slopes along an arithmetic scale. Comparing the frequency distributions, they conclude, "Un-

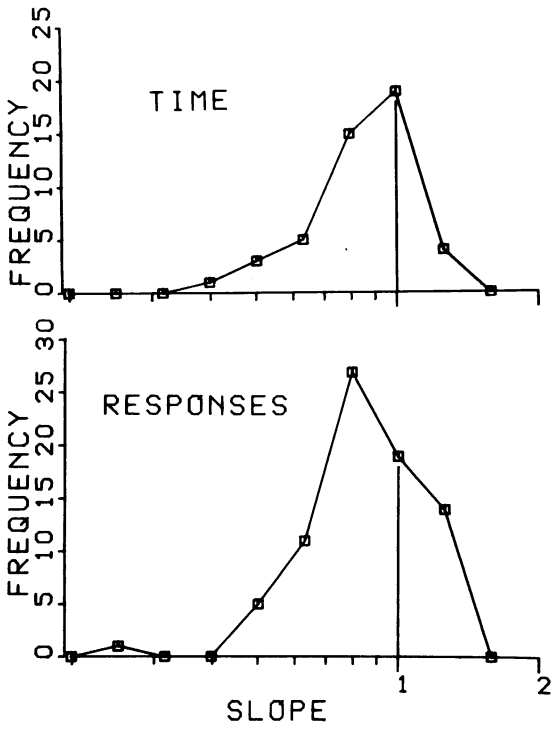

Fig. 1. Frequency distributions of the slope a obtained by fitting the power law of choice to the various sets of data compiled by Wearden and Burgess (1982). Vertical lines show the location of 1.0. The upper distribution shows slopes fitted to time ratios; the lower distribution, to response ratios. Note logarithmic axes for slope.

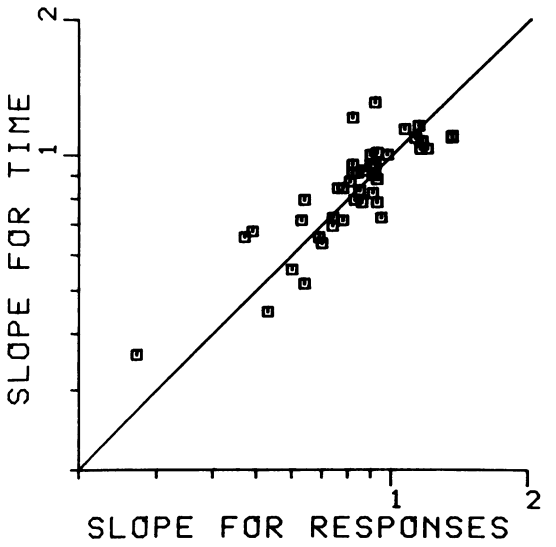

Fig. 2. Scatter plot of slope fitted to time ratios versus slope fitted to response ratios, for all experiments, compiled by Wearden and Burgess (1982), in which both measurements were made. Diagonal line shows locus of equality. Note logarithmic axes.

dermatching (slope less than 1.0) was predominant in both response-distribution and time-allocation measures" (p. 341).

Figure 1 shows frequency distributions of the slopes gathered by Wearden and Burgess represented on logarithmic scales. The class intervals used for Figure 1 were .1 log unit wide as in my 1979 paper (incorrectly reported there as .2 log unit). The time slopes clearly show a single strong mode in the class interval including 1.0, which contained slopes from .9 to 1.12. The response slopes, in contrast, show a single mode at about .8—undermatching. In other words, when slope is scaled logarithmically, the slopes gathered by Wearden and Burgess agree completely with my earlier analysis (Baum, 1979, Figure 3).

Comparing time slope with response slope in those experiments where both were determined, Wearden and Burgess examine differences between the slopes, instead of ratios. As a result, they draw several erroneous conclusions. Figure 2 shows that, when the slopes are represented on logarithmic scales and compared, the results agree with the results of my analysis: Inequality occurs commonly, and when the slopes differ substantially, the time slope is always the larger (Baum, 1979, Figure 4). Too little overmatching occurs for any firm conclusions about its relative frequency.

Instead of more discussion of modes and variation in slopes from study to study, we need experiments that give some understanding as to why time slopes generally exceed response slopes and what factors cause the slopes to vary. Davison and associates report evidence that scheduling of reinforcement can increase or decrease the choice slope (Davison, 1982; Taylor & Davison, 1983). Research on changeover requirements indicates that increasing the cost of changeover can increase the choice slope across the entire range from undermatching to overmatching (see Baum [1982] for a summary, and Dunn [1982]). Keller and Gollub (1977) found that extended training can change performance from matching to undermatching. More research along such lines will allow us to decide whether the matching law can be upheld or whether it must be replaced.

RELIABILITY

The second claim of Mullins et al., that I used incorrect methods of evaluating the fits of the various data sets to the matching law, derives from the narrowness of their interpretation of the r^2 statistic and an unfounded preference for parametric statistics. The statistic r^2, often called the "proportion of variance accounted for," equals

$$1-\frac{v_e}{v},$$

where v is the total variance in the data, and v_e is the variance around the fitted curve or line. In usual linear regression, in which both slope and intercept are allowed to vary, r^2 varies between 0 and 1.0, because v_e cannot exceed v. One way to assess how far a set of data deviates from matching is to ask how much the fit to the data deteriorates if one assumes matching. This

means fitting a line with slope of 1.0; only the intercept is allowed to vary. Unless the fitted slope happens to equal exactly 1.0, the two-parameter r^2 (H) will exceed the one-parameter r^2 (H'). In addition, H' may even decrease to less than 0, because v_e, the variance around the line with assumed slope 1.0, may exceed v, the total variance in the data. Mullins et al. state, "It is sufficient to point out that H' may be negative to undermine the reliability of the measure" (p. 325). They are mistaken, because although r^2 is called the "proportion of variance accounted for," it is really a comparison between the two variances v and v_e. Negative H' has a straightforward meaning. It means that the fit to the line with slope 1.0 is so much worse than the two-parameter fit that one would actually have done better to assume a slope of zero. This occurs when the fitted slope is less than .5.

Mullins et al. go on to suggest a parametric statistical test for deviation from a slope of 1.0. For reasons that remain unclear, they prefer it to the nonparametric test that I used. They state, "If it can be assumed that the distribution of the deviations about the regression line is approximately Normal (as could certainly be done for many of the data sets in question), a parametric test . . . would be more appropriate than a nonparametric test" (p. 326). Rarely, probably never, will an experiment produce enough data to permit verifying the assumption. Precisely this *un*certainty led me to prefer the nonparametric test. I cannot see how a test requiring an unverifiable assumption can be more "appropriate" than one that makes no such assumption.

ESTIMATION

The third claim of Mullins et al., that I was mistaken in suggesting the range of slopes from .9 to 1.11 to approximate matching, points up a philosophical difference between us. In suggesting the range of slopes, I was reporting my findings. I had no special concern for the question, "How far can a slope differ from 1.0 and still be consistent with matching?" That depends on the degree of unsystematic variation in the data. I was trying, rather, to answer the question, "How close can a slope be to 1.0 and be considered a reasonable approximation to 1.0?" Someone else might answer with a smaller or larger range than I did, but Mullins et al. seem to feel the question itself is at fault. They argue that each data set should be submitted to statistical test, even those that are fitted by slopes close to 1.0. They give as an example, "if all of a large number of observations lie exactly on a straight line of, for example, slope .97 it would hardly be appropriate to decide that the 'true' slope is 1.0" (p. 326). In practice, of course, such an occurrence is extremely unlikely. But, suppose one experiment produced a slope of .97 that by some statistical test was significantly different from 1.0 with a p of .001. Mullins et al. would insist that the slope was different from 1.0. I would never wish to take such a position. No one will ever repeat the experiment exactly. Even if the test were perfectly valid (are all assumptions ever met?), why isn't this the one case in a thousand that p tells about? If I find a set of data fitted by a slope of .97, I will always want to call that a good approximation to 1.0. No statistical test will ever substitute for plain common sense.

William M. Baum
University of New Hampshire

REFERENCES

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

Baum, W. M. Choice, changeover, and travel. *Journal of the Experimental Analysis of Behavior,* 1982, **38,** 35-49.

Davison, M. Preference in concurrent variable-interval fixed-ratio schedules. *Journal of the Experimental Analysis of Behavior,* 1982, **37,** 81-96.

Dunn, R. M. Choice, relative reinforcer duration, and the changeover ratio. *Journal of the Experimental Analysis of Behavior,* 1982, **38,** 313-319.

Keller, J. V., & Gollub, L. R. Duration and rate of reinforcement as determinants of concurrent responding. *Journal of the Experimental Analysis of Behavior,* 1977, **28,** 145-153.

Mullins, E., Agunwamba, C. C., & Donohoe, A. J. On the analysis of studies of choice. *Journal of the Experimental Analysis of Behavior,* 1982, **37,** 323-327.

Taylor, R., & Davison, M. Sensitivity to reinforcement in concurrent arithmetic and exponential schedules. *Journal of the Experimental Analysis of Behavior,* 1983, **39,** 191-198.

Wearden, J. H., & Burgess, I. S. Matching since Baum (1979). *Journal of the Experimental Analysis of Behavior,* 1982, **38,** 339-348.

Received June 30, 1982
Final acceptance December 22, 1982