Two subjective scales of number

WILLIAM P. BANKS and MARK J. COLEMAN
Pomona College, Claremont, California 91711

The way in which the apparent magnitude of numbers grows with their absolute magnitude was measured with a modified version of the direct technique Marks and Slawson (1966) used to determine the psychophysical exponent for loudness. This modified technique required subjects to estimate how evenly and randomly a sequence of integers appeared to sample the numerical continuum. The results indicate that the apparent magnitude of numbers increases with a decelerated power function of their arithmetic magnitude when a series samples from an open-ended range. However, when an upper boundary of the range is specified, the subjective scale seems to be linear. Random productions of numbers parallel the results found with judgments of presented sequences. The two scales of number provide the basis for an interpretation of the difference between magnitude and category scales: that subjects use numbers differently when the response scale is open-ended (magnitude estimation) than when it has a fixed upper limit (category scale). Given the assumption that subjects use numbers in this way in the two tasks, the qualitative relation between magnitude and category scales is predicted.

Since Garner, Hake, and Eriksen (1956) discussed the effect response systems can have on reports of perceptual events and sensory magnitudes, there have been a number of attempts to determine the effects of response scales on the results of direct-scaling techniques. In the magnitude estimation paradigm, the most popular and versatile of the direct-scaling techniques, the response scale is the numerical continuum, and interest in response effects in direct scaling has naturally focused on numerical responses. Attneave (1962) provided a stimulating discussion of the issues relating to use of numerical responses, and, in the same paper, he presented probably the first reported attempt to scale the numerical continuum. From the results of this informal experiment, Attneave concluded that the subjective magnitude of numbers grows with the .4 power of their absolute magnitude. If subjects in magnitude estimation tasks used such a compressed scale of number in selecting their responses, a stimulus called "10" would be subjectively only two or three times as large as a stimulus called "1," not 10 times as great.

A few studies subsequent to Attneave's 1962 paper have yielded subjective scales of number, but it has been only a few because the numerical continuum presents a very difficult problem. Direct scaling techniques such as a ratio production or magnitude estimation cannot be used because subjects can use arithmetic algorithms to estimate the "correct" answer in a way that is impossible with psychophysical continua. More important, magnitude estimation applied to the scaling of number would necessarily be subject to the distortions of the response system, and a direct scale of number would therefore be contaminated by the very factors it is intended to measure. Indirect techniques, at least the classical ones, also seem inapplicable to the numerical continuum. For example, a JND scale of integers is absurd on the face of it: Two numbers are either equal or they are not (however, Weissman, Hollingsworth, & Baird, 1975, have obtained psychometric functions for differences between *sets* of numbers).

In spite of the difficulties of scaling the numbers (or possibly because of the difficulties), at least three difficult forms of the subjective scale of the numerical continuum have emerged from scaling studies: (1) a linear scale (Goude, 1962; Rosner, 1965); (2) a logarithmic scale (Ekman, 1964; Ekman & Hosman, 1965; Moyer & Landauer, 1967; Rule, 1969); and (3) a decelerated power function scale (Curtis, Attneave, & Harrington, 1968; Curtis & Fox, 1969; Schneider, Parker, Ostrosky, Stein, & Kanow, 1974). Banks and Hill (1974) used several scaling techniques and were able to conclude that the scale is compressive, but they found that either a logarithmic or a power function (exponent of .6) fit their results equally well. Banks, Fujii, and Kayra-Stuart (1976) concluded that a compressed scale of number is needed to account for the pattern of reaction times for comparative judgments among pairs of digits. (Moyer & Landauer, 1967, came to the same conclusion.) The Banks et al. (1976) study was, however, similar to that of Banks and Hill (1974) in its inability to distinguish between a compressed power function and a logarithmic function.

95

Experiments 1-5 in the present study attempt to decide among these proposed scales by use of a variation of the technique Marks and Slawson (1966) devised to make a direct test of the S. S. Stevens (1956) power function for loudness (cf. also Borg, Edstrom, & Marklund, Note 1, for another variation of the Marks and Slawson techniques). Marks and Slawson presented their subjects with 2.5-sec bursts of auditory pulses that increased in loudness according to various functions during presentation, and subjects rated the apparent deviation from linearity of the growth in loudness of each burst. The functions by which loudness increased were chosen in such a way that they would each provide a linear increase over a different mapping of sound pressure level onto the subjective continuum. The psychophysical law whose growth function was rated overall the least nonlinear was taken as the best psychophysical law for loudness. The results indicated a power-law psychophysical function with an exponent of .6. This is the same psychophysical function for sound pressure level as is obtained from numerical magnitude estimations of loudness.

The Marks and Slawson technique cannot be used in unmodified form to estimate the subjective scale for number. Most subjects would apply a mathematical definition of linearity (i.e., $N_j - N_{j-1} = k$) to numbers in any monotonically ascending or descending series, and the series judged most linear would automatically be an arithmetic one, regardless of the subjective scale of number. Consequently, the task of Experiments 1-5 was to judge how evenly and randomly a sequence of irregularly ordered numbers samples the numerical continuum. The logic of the Marks and Slawson technique is little altered by the present variation, and they could have performed their experiment in a manner analogous to the present one. In such an experiment, they would have presented subjects with a series of pulses varying unsystematically in loudness. The loudness of the pulses would be selected by sampling rectilinearly over various mappings of sound pressure level. The series that used the "true" psychophysical mapping should sound most random. Sampling according to other mapping functions would result in series that would seem to have too many loud pulses or too many soft ones.

The first five experiments led us to the unexpected conclusion that the scale of number depends on the range over which the numbers are sampled. If the range has finite limits at both ends, subjects find rectilinear sampling to be the most uniform and random. If the top end is unbounded, subjects find a decelerated power function to be the best. There seem, therefore, to be two subjective scales of number, one for closed ranges and one for infinite ranges. Experiment 6 shows that production of random numbers shows the same relationship: rectilinear sampling over closed ranges and compressive sampling

(more small than large numbers) over open ranges. These results suggest that people have available two separate scales of numbers, and the range of numbers determines which scale is used. If one scale is used in magnitude estimation and the other in category judgment, many relations between these two scaling techniques are explained.

EXPERIMENT 1

Method

Subjects. Twelve male and female Pomona College undergraduates volunteered to be unpaid subjects. The subjects all had had previous experience in random number generation experiments similar to those of Banks and Hill (1974).

Materials. Four sequences of 10 integers were generated in the range of approximately 1 to 2,000. The numbers were sampled rectilinearly from "rulers" graduated in logarithmic, linear, inverse square (exponent 1/2), and inverse cube (exponent 1/3) units. The linear set had 1 as its smallest number, and every 225th number above for the next nine. Small integral amounts were added to or subtracted from each number to conceal the underlying orderliness of the numbers. The random numbers from the logarithmic scale were generated by means of the function $N = e^{kn}$, where k took on the values, 1, 2, ... 10, and n was chosen such that when $k = 1$, the function was close to 1.0, and when $k = 10$, the function was close to 2,000. The random numbers from the inverse square and inverse cube scales were generated with the function $N = (ka)^E$, where k took on the values 1, 2, ... 10, E was 2 for the inverse square ruler and 3 for the inverse cube ruler, and a was chosen such that $N = 1$ for $k = 1$, and $N = 2,000$ for $k = 10$ for each ruler. Values of N for the nonlinear scales were rounded off to the nearest integer, and the numbers for all four functions were put in the same irregular order (the sequence of ordinal magnitudes in all four irregular orders was the same). The four sequences of scrambled numbers were then tape-recorded by a male speaker.

Procedure. The subjects had the following instructions read to them: "You are going to hear four series of random whole numbers played to you on the tape recorder. After hearing each series I would like you to rate how randomly and evenly it seems to you to sample numbers. Use a 1-10 scale for your ratings, where 1 means least random and 10 means most random." Questions were resolved by repeating the instructions or, failing that, by stating that randomness implied that all numbers within the range sampled had an equal chance of being selected. Once the subjects indicated satisfaction with the instructions, the four sequences were played and then judged individually immediately after being played. The four sequences were administered for judgment in an order counterbalanced for position in the session over subjects.

Results

Figure 1 shows the mean judgments of randomness of the numbers derived from each series. The linear and logarithmic series were judged least random, and the inverse cube function stands out as considerably more random than the others. Only two subjects rated either the linear or logarithmic series more random than the cube series. The four means are reliably different at p < .01, with an F(3,36) of 10.98. The four medians differ at the same level according to the Friedman test [$\chi^2(3) = 171.38$].

Discussion

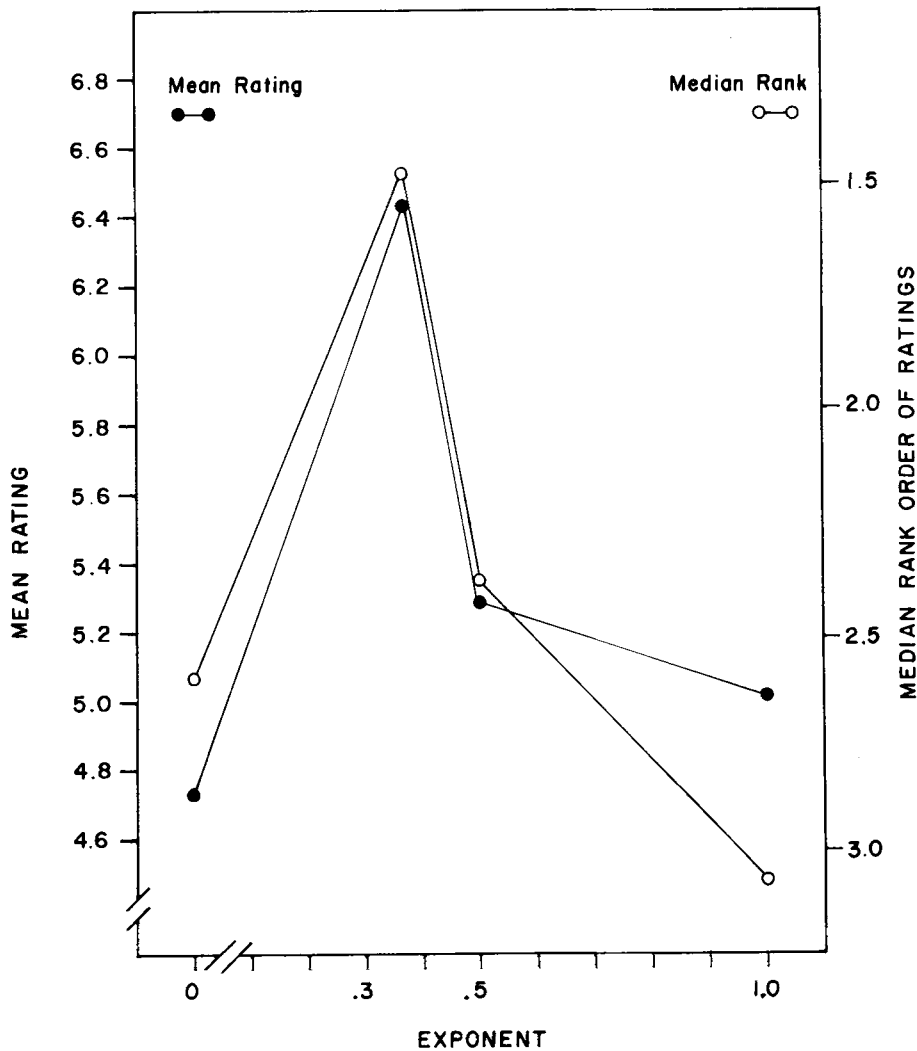As is clear from Figure 1, the results indicate that

Figure 1. Results of Experiment 1: Mean rating and median rank of rating of subjective random-ness of randomized series of integers selected according to a rectilinear sampling rule from numeri-cal scales compressed according to power functions with exponents of .3, .5, or 1.0. A fourth com-pression, logarithmic, is plotted as though it were a power function with an exponent of .0.

the subjective scale of number is decelerated. A power function with an exponent less than 1.0 seems to be the best-fitting function, but it is not possible to estimate the size of the exponent confidently from the present results because only two decelerated power function scales were investigated. At the very least, it can be stated that the scale is compressive, but not so com-pressive as a logarithmic function. However, if the trend of the results as seen in Figure 1 is approximated with a smooth curve, it seems to indicate more specif-ically that the psychophysical power function for number is one with an exponent less than .5. An ex-ponent in this range would be in approximate agree-ment with Attneave's (1962) estimate and with the best exponent in the power law tried by Banks et al. (1976). Such a small exponent is, on the other hand, too small

to agree with the estimates of Curtis and his colleagues (Curtis et al., 1968; Curtis & Fox, 1969) or with the estimates of Banks and Hill (1974). In both cases, estimates fell in the range from .5 to 1.0, which was not even examined here.

Experiment 2 samples decelerated exponents more systematically and completely than those in Experi-ment 1 to give a better estimate of the exponent for number. Experiment 1 does not locate the peak of the estimating function very precisely, nor does it eliminate the possibility of a second peak in the range .5 to 1.0.

EXPERIMENT 2

Method
Subjects. Twelve male and female Pomona College undergrad-uates were recruited to serve as unpaid subjects.

Materials. Eight series of 15 integers were generated in the range from approximately 1 to 100,000. Seven of the series were generated using power functions as in Experiment 1 and with exponents equal to .1, .3, .36, .5, .7, .9, and 1.1. The third exponent, .36, was chosen to create a series approximating the random numbers generated by an independent group of 10 subjects in a random number generation task similar to Banks and Hill's (1974). The exponent comes from a least-squares fit to the order statistics for the first 25 numbers produced. The largest number (100,000) in each of these seven series was not presented in order to avoid an obvious upper limit. In addition to these, an eighth series was presented. This eighth series was produced by the same independent group of 10 subjects whose number scale for random production had an exponent of .36. Rather than fitting a curve to obtain this series, we used the means of the numbers subjects gave as productions. The first 14 numbers given by these subjects were put in order from smallest to largest, and the geometric mean at each ordinal position was taken across subjects. (These means constitute the first 14 order statistics, as discussed by Banks & Hill, 1974). These means ranged from 1.4 to 9,973. In order to give the series a maximum of about 100,000, each geometric mean of each ordered position was multiplied by 10. In this series, when only the first 14 numbers are used, the exponent fit to the order statistics is .39.

The order of the numbers within each series and the order of presentation of the series were then randomized and tape-recorded by a male speaker.

Procedure. The procedure was the same as that used in Experiment 1.

Results

These results, shown in Figure 2, generally agree with those obtained in Experiment 1. The two series based on subjects' productions, plotted as exponents of .36 and .39, were judged to be most random. These two were not judged to be significantly different from each other according to a liberal least significant difference test ($\alpha = .50$, LSD $= .66$). As the exponent increased in value, the mean randomness rating fell, with the exception of the exponent .9, for which a slight increase, followed by a decrease, was measured. The mean ratings were reliably different at $p < .01$ [$F(7,77) = 3.10$].
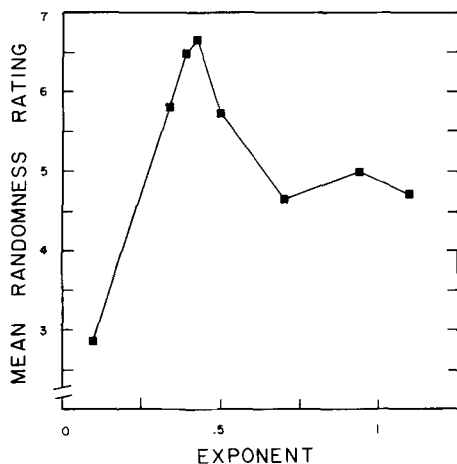

Figure 2. Results of Experiment 2: mean ratings of randomness of series generated analogously to those in Experiment 1. See text for details of generation procedure.

The function in the figure seems to indicate clearly that the most random power function series is generated with an exponent between .3 and .4, just as Experiment 1 found. Experiment 3 was conducted because it seemed necessary to replicate the findings of Experiment 2 with samples taken from smaller numbers than before. Here the range 1-2,000 was used, instead of 1-100,000 as in Experiment 2. The range 1-2,000 more nearly approximates the range of numbers people use when asked to make judgments in magnitude estimation, and it probably samples numbers people "know" better (or at least use more frequently) than most of those in the 1-100,000 range. The small exponent estimated from samples from the 1-100,000 range may be overly influenced by the very compressed subjective scale for the very large numbers included. Banks and Hill (1974) found the subjective scale for number to become progressively more compressed for larger numbers. The exponent of .6 fit the range from 1 to 1,000 quite well, but above 10,000, the compression progressively approached a logarithmic scale as the numbers grew larger.

Another reason for repeating Experiment 2 with a narrower range is that the presence of large numbers in samples may have made them seem less random. Subjects may assume that large numbers should almost never be selected and that their presence in the sample indicates nonrandomness. Since the proportion of large numbers declines as the exponent decreases, the assumption that large numbers indicate nonrandomness will create a bias toward smaller exponents. Use of a sample from a range restricted to 2,000 or less will test for such a bias. There should be less of a bias toward small exponents in such a case, and the randomness judgments should shift to series with larger exponents if the bias against large numbers were indeed a factor.

EXPERIMENT 3

Method

Subjects. Nine male and female Claremont colleges undergraduates served as unpaid subjects for the present experiment.

Materials. Seven sequences of 15 integers were generated in the range from approximately 1 to 2,000. All were generated according to the power function of Experiments 1 and 2, with exponents of .1, .3, .36, .5, .7, .9, and 1.1. The 15th integer in each series (2,000) was omitted from presentation to avoid that number's serving as an obvious upper limit. The first 14 integers in each series were put in random order as before.

Procedure. The presentation of the stimuli was identical to that used in previous experiments.

Results

The exponent .36 produced the highest randomness ratings, with a decrease in ratings with exponents smaller and larger in value. This finding is demonstrated in Figure 3 with a plot of exponent value vs. mean randomness rating. Analysis of variance shows the randomness ratings significantly different at $p < .0001$ [$F(6,48) = 6.73$].
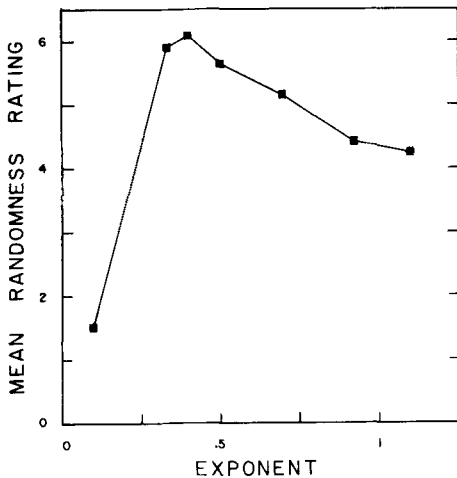
Figure 3. Results of Experiment 3: mean ratings of randomness of series generated by rules similar to those used for power-function series in Experiments 1 and 2. However, in this case, no numbers above 2,000 were used in the series.

The results of Experiment 3 confirm those of Experiments 1 and 2: A compressive power function with an exponent between .3 and .4 produces the series judged most random. Neither the requirement to generate before judging, as in Experiment 1, nor the wide range of Experiment 2 was responsible for the previous finding. The following two experiments were conducted in an attempt to determine under what conditions subjects would judge the randomness of series linearly, that is, according to the criterion of equiprobability over a linear scale. Given a relatively small number of alternatives, subjects can judge equiprobability quite accurately (Peterson & Beach, 1967, pp. 30-31), and it seems they should be able to do so under some conditions here. Experiments 4 and 5 use series that are quite clearly from a *closed* scale, in contrast with the *open* scale of Experiments 1-3. The impression of an open scale came about because those experiments omitted the highest number, which was either 2,000 or 100,000, depending on the experiment. Thus, the largest number varied from series to series. In Experiments 4 and 5, the endpoint of the series was 1,000, and it was always presented with the other numbers to give a consistent upper anchor to the series.

EXPERIMENT 4

Method

Subjects. Twenty-nine Claremont colleges undergraduates were recruited from two psychology courses to be unpaid subjects.

Materials. Five sequences of 10 integers were generated in the range from approximately 1 to 1,000. The numbers were rectilinearly sampled from "rulers" graduated in units with exponents of .1, .3, .5, .7, and .9, according to the same method used in Experiments 1-3. The function $N = (ka)^E$ describes the different scales used, with E equaling one of the five exponents, k taking

on the integral values 1-10 inclusive, and a chosen for each series to give a maximum close to 1,000. Values of N were rounded off to the nearest integer, resulting in a minimum of 0 and maximum of 1,000 in each series.

In the previous experiments, the largest number in the sequence (the "de facto" upper limit) varied from series to series, producing (we assume) the impression that there was no fixed upper limit to the range being sampled. In these series, on the other hand, the largest number was 1,000 and the smallest was 0. Subjects may have judged the first series they heard as though it were sampled without an upper limit, but, by the second series, they should have observed the consistent limits and judged accordingly from then on. Any systematic effort of a shift in strategy should be equally distributed over series because position of series in the order of presentation was completely counterbalanced over subjects.

Procedure. The procedure was identical to that used in the previous experiments, except that the instructions ended with the following: "Do not attempt to apply any mathematical definition of randomness. Just judge on the basis of an intuitive feel for the randomness with which the numbers are sampled. Use a 1-10 scale for your ratings, where 1 means least random and 10 means most random."

Results and Discussion

Figure 4 plots the mean randomness ratings as a function of exponent. The series generated with the exponent .1 was seen as least random, and the series generated with the exponent .9 as most random. The randomness ratings increased as the exponent approached the value of 1. An analysis of variance revealed a significant difference between the mean randomness ratings [$F(4,112) = 13.96$, $p < .0001$].

The results show a function somewhat different from that found in the previous experiments. Here we find greater subjective randomness as the exponent approaches unity. The results suggest that an exponent of 1.0 (linear scale) or even an accelerated scale might turn out to be best. Experiment 5 replicated
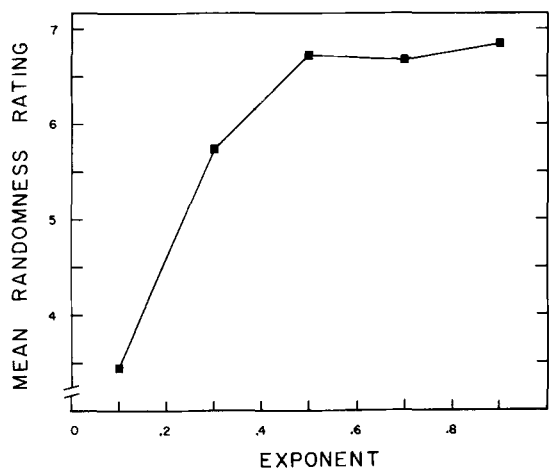

Figure 4. Results of Experiment 4: mean ratings of randomness of series generated by power functions as in the previous experiments, but, in this case, an explicit upper limit of approximately 100 was used so that subjects knew that samples were limited to 1-, 2-, and 3-digit numbers.

Experiment 4 with two changes. First, subjects generated random numbers before judging the series in order to test whether generation imposes a bias toward nonlinear series. All of the subjects in Experiment 1, and some of those in Experiments 2 and 3, had been given a random number generation task prior to judging the series. Second, an exponent of 1.1 was added to test whether the series with an exponent of 1.0 was a maximum.

EXPERIMENT 5

Method
Subjects. Ten male and female Pomona College undergraduates served as unpaid subjects.

Materials. Six series of 10 integers were generated with the equation described in Experiment 4. The exponents were .1, .3, .5, .7, .9, and 1.1.

Procedure. All subjects were given the task of random number generation before the randomness estimation task. Instructions similar to those used by Banks and Hill (1974) were read to the subjects. Subjects gave 25 numbers, and all productions were recorded on audiotape. Immediately after random number generation, subjects received instructions essentially the same as those for Experiment 4. The six sequences were again counterbalanced for position over subjects, and all subjects were tested individually.

Results and Discussion
The subjects' random numbers under 10^6 were analyzed by the same techniques used by Banks and Hill (1974): Each subject's random productions were put in rank order, and the geometric mean for each ordinal position was taken across subjects. The result was a measure of central tendency for each ordinal position, known as the order statistic. These numbers were, incidentally, those used in Experiment 2 for the series with the exponents of .36 and .39.

In producing random numbers, subjects selected more smaller numbers than they did larger ones. More precisely, in agreement with the Banks and Hill analysis, subjects produced the numbers as though selecting from a ruler compressed according to a power function with an exponent less than 1.0. The power function fit to the numbers these subjects gave has an exponent of .36 if the entire set of 25 numbers is used, and an exponent of .39 if only the first 14 are used. These figures seem in some conflict with the Banks and Hill estimate of .6 for the power function for number, but the conflict is not as great as it might seem. Banks and Hill's estimate of the exponent for the subjective magnitude of numbers is not based simply on the fit of a power function to the full range of productions, but also on the fit to smaller numbers, which are used most often in direct scaling experiments. Fitting a power function to the numbers Banks and Hill obtained under 1,000 alone in their Experiment 2 gives an exponent of .49. When only the numbers under 1,000 in the present productions are considered, the exponent for the power-function fit is .52.
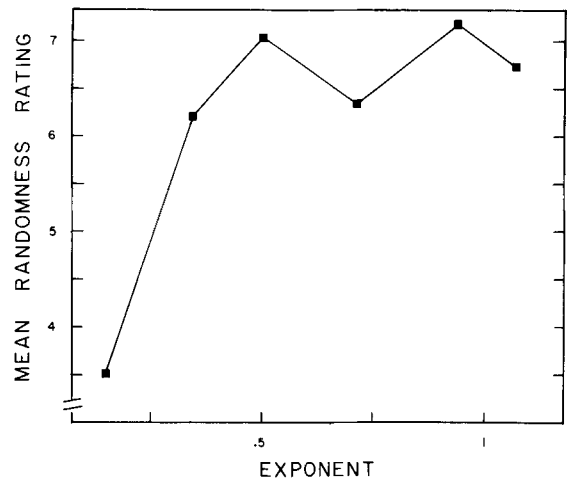

Figure 5. Results of Experiment 5: mean ratings of randomness of series generated by power functions as in the previous experiments, and with an explicit upper limit as in Experiment 4. Here, however, subjects generated random numbers before judging the series in order to allow a test of whether the difference between the functions of Experiment 4 and Experiments 1-3 resulted from experience subjects had had in the earlier experiments with random number generation before estimation.

Figure 5 plots the results of the second half of the experiment: randomness rating as a function of exponent. This replicates Experiment 4, with [F(5,45) = 9.52, p < .0001] for the difference among the series' means. The randomness ratings increased as the exponent approached unity, and fell of at 1.1. The results of Experiment 5 leave little question that subjects judge randomness according to a linear function when numbers are sampled from a range specified at both ends. The differences between Experiments 1-3, which revealed a compressive subjective scale of number, and Experiments 4 and 5, which revealed a linear scale, seem to depend on whether the range of numbers given to the subjects has a definite upper limit.

EXPERIMENT 6

Experiments 1-5 show a great deal of similarity between number production and number perception with an open-ended range. The productions reported in Banks and Hill (1974), as well as in this paper, imply a decelerated power function scale of number, and the present judgments of randomness in an open-ended scale imply that the subjectively most random and "even" sampling of numbers also comes from a compressed scale. The precise sizes of the exponents of power functions for the subjective scale in the two cases are even approximately the same. Furthermore, the series judged subjectively to be the most random of all is composed of the order statistics of numbers subjects actually produced. Our experiments allow us

to compare production and estimation for a scale without an explicit top end, but we have only estimation data for series with closed ends and no productions. Experiment 6 examines whether random productions in a restricted range imply a linear scale of number, as the judgments do. If the linear scale found for the judgments in the closed scale reflects the subjective scale of number and not some strategy subjects use to estimate randomness in this special case, then the same scale should be found in productions in a closed scale.

Method

Subjects. Twenty-one subjects, male and female students at Claremont colleges, volunteered to perform in this experiment while waiting to serve in another, unrelated one. All were naive to this experimental task.

Procedure. Subjects first gave 20 to 25 numbers under unconstrained random number production instructions identical to those used by Banks and Hill (1974), Experiments 1-3. They were then asked to produce random integers as before but to constrain their productions to a certain range. Three different ranges were given, each to 7 of the 21 subjects. Subjects were assigned to one of the three range groups cyclically on the basis of their order of volunteering for the experiment, with the first subject in Group 1, the second in Group 2, third in Group 3, fourth in Group 1, and so on. The range for Group 1 was the integers 1-9, for Group 2, the integers 10-99, and for Group 3, the integers 100-999. Each subject was allowed to produce at least 10 integers before being stopped. All productions were tape-recorded. The first 10 numbers produced in the experimentally defined range were transcribed for analysis.

Results

Subjects' responses were put in rank order, and the geometric mean for each position was computed across subjects for each of the three constrained conditions, as well as for the open-ended condition. This technique of analyzing the production yields the order statistics (cf. Banks & Hill, 1974) for the series.

The order statistics for the unconstrained series gave subjective scales very similar to those obtained by Banks and Hill (1974). The subjects sampled more small than large numbers in any range of their productions examined. The subjective scale for the unconstrained number (looking only at the first 20 numbers subjects gave) was fit by a power function with an exponent of .55. The r for the fit was .92. On the other hand, the order statistics for the restricted ranges were well fit by straight lines using the least-squares method: for the range 1-10, $r^2 = .995$; for the range 10-99, $r^2 = .974$; for the range 100-999, $r^2 = .990$. When the responses for the range 10-99 are divided by 10 and responses for the range 100-999 are divided by 100, and the data from all three ranges subjected to an analysis of variance, there are no reliable differences between the functions at all. It is as though subjects in these three conditions adjusted output perfectly to the ranges they had and sampled rectilinearly in each case. The linearity of sampling within restricted intervals replicates the similar results obtained by Noma and Baird (1975). Figure 6 shows
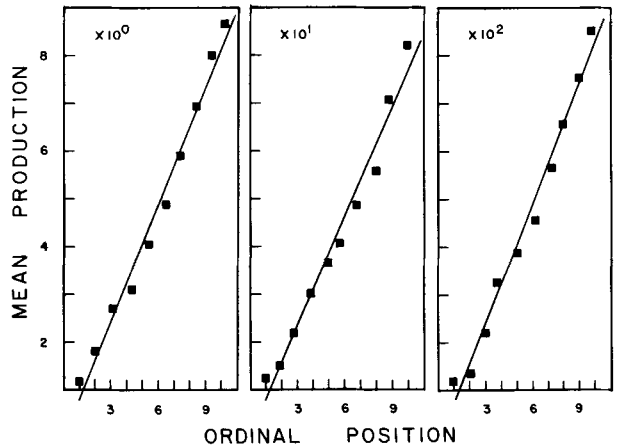

Figure 6. Results of Experiment 6: mean of order statistics (see text) as a function of their rank order for integers generated randomly by subjects in the range 1-9 (left panel), 10-99 (center panel), and 100-999 (right panel). These functions have been scaled to fit in the same space. The factor needed for rescaling both axes is inserted in the upper left-hand corner of each panel.

how the mean of the order statistics increases with their rank. When the ranges are appropriately scaled, the similarity (and the linearity) of the functions is remarkable.

GENERAL DISCUSSION

Experiments 1-3 show a sort of behavior regarding numbers very different from that exhibited in Experiments 4-6. The generalization that seems dictated by our results is that subjects perceive the numerical continuum as compressive when they deal with a numerical scale that is open at the upper end, and they perceive the numbers linearly when given a range limited at both ends. This generalization holds whether subjects are judging the apparent randomness of numbers sampled from open and closed ranges or whether they are generating subjectively random samples of integers from limited or unlimited ranges.

While this generalization seems amply supported by our results, it is nonetheless ambiguous on whether the shifts in compression come from changes in scale compression or sampling rule. The choice of compressive scales as most random in Experiments 1-3 could have been made either because subjects operate with a compressed subjective representation of number or because they apply a nonrectangular sampling rule to a perfectly linear representation of the numerical scale. In fact, there are not just these two alternatives, but a whole class of them, composed of the set of possible scale compressions combined with whatever sampling distribution is necessary in each case to complement the degree of scale compression and produce the functions obtained from the subjects. By the same reasoning, the linear functions obtained in Experiments 4-6 could result from a linear numerical

scale combined with a linear sampling function or from any of an infinite set of combinations of nonlinear numerical scales and nonlinear sampling rules.

Banks and Hill (1974, Experiment 4) provided a test of whether the sampling rule or the subjective numerical continuum was compressive in unlimited-range random number generation. In this test, subjects were shown a drawing of a "number line" that was to represent the numerical continuum, going from small numbers at the left end to large numbers at the right, which terminated in a rightward-pointing arrowhead to indicate there was to be no explicit upper limit to the numbers. A single vertical histogram bar was drawn on the number line near either the left or the right end of the line. Subjects had previously generated numbers according to instructions with no constraint on the upper limit, and were told that this bar represented their instructions for the next random generation task. When the bar was at the left end of the scale, they were to generate numbers within an interval of their own choosing among the smaller numbers; when the bar was at the right end, they were to generate within an interval among the larger numbers. The definition of the width of the range sampled and of the meaning of "larger" and "smaller" numbers was left to the subjects. The width of range they chose was, in fact, the dependent measure. The results showed that subjects used a wider numerical range as the range was placed higher on the continuum, despite the fact that for both high and low placements the bars were drawn to the same width. Such an outcome implies that subjects assumed that the same bar width spanned a greater numerical range for larger numbers and thus that the continuum was compressed, not the sampling rule.

In Experiment 4 of Banks and Hill (1974) the number line was explicitly defined as open-ended. The result is therefore in agreement with the present set of studies, which show that a numerical scale unbounded at the upper end is perceived as compressed. A further prediction from the present studies is that the continuum would be perceived as linear, or at least as less compressed, when the upper end is bounded. To give an informal test of this prediction, we asked nine subjects waiting for other experiments (who had never performed in a number experiment and did not know the purpose of this one) to generate numbers in intervals on a "number line" like that used in Experiment 4 of Banks and Hill. However, they had both an unbounded range, like that of the previous Banks and Hill experiment, and a bounded range. In both the unbounded and bounded ranges, the numeral "1" was written just above the boundary marker at the bottom end of the range. Above the upper boundary marker at the top end of the bounded range, "1,000" was written to give a numerical upper limit. There was nothing written over the arrowhead at the right end of the unbounded range.

When shown the unbounded number line, the mean range of numbers subjects gave was 178 for the interval at the low end of the scale and 964 for the interval at the high end of the scale. This difference is reliable, with $z = 3.23$, $p < .01$. This is actually a stronger effect than previously found, where the mean range was 45 and 82 for small and large numbers, respectively, but there were a few procedural differences between this and the previous experiments and no reason to expect exact replication. The bounded number line, as predicted, yielded a roughly equal mean range for intervals among both small and large numbers, with 114 and 131 for small and large, respectively, and the difference falls far short of reliability. This experiment therefore supports the hypothesis that the differences between open and closed ranges in both judgment and production of subjectively random sequences of numbers derive from differences in scale compression in the two cases rather than from differences in sampling rule.

If there are two different subjective scales of number, as our results seem to show, then we can draw some conclusions regarding the difference between results of two widely used scaling techniques that require subjects to give direct numerical estimates of stimulus magnitudes. These two techniques are magnitude estimation and category judgment scaling. In both techniques, subjects report the position of a stimulus on a unidimensional scale, and in both, subjects give numerical responses to indicate the scale value of each stimulus. The numbers subjects give are averaged to give the mean scale value for each stimulus. The two techniques differ in the constraints on the subject's numerical responses. In category scaling, subjects are given a small set of categories that are usually denoted by integers, and they can only use this set of integers as responses. In magnitude estimation, on the other hand, subjects' numbers are intended to express the apparent ratio of magnitude of experimentally presented stimuli to an explicit or implicit reference magnitude. In magnitude estimation, the responses are relatively unconstrained compared with category scaling. Subjects are not restricted to integers, and they need observe no upper limit to their responses.

Similar as they may seem, category and magnitude tasks have some very different characteristics. First, it appears that category scales are more affected by stimulus spacing and procedural variables than magnitude estimation scales are (cf. Marks, 1968; J. C. Stevens, 1958; S. S. Stevens, 1971; Stevens & Galanter, 1957). Second, despite the variations in the form of the category scale that may be induced by variations in technique, the category scale values for a continuum are generally a curvilinear function of the magnitude estimation scale values for the same continuum. The form of this relation is such that when category values are plotted as a function of magnitude estimation, the function is typically a concave downward, nega-

tively accelerated one, with category scale values moving relatively more rapidly than magnitude values at the bottom of the scale and less rapidly at the top (Eisler, 1963; J. C. Stevens, 1958; S. S. Stevens, 1971).

Explanations of the difference between the scales have usually been based on the hypothesis that different operations of judgment are induced by the instructions in the two cases. However, another difference between the two techniques that has been relatively neglected is the range of responses allowed the subjects. Magnitude estimation allows subjects an open-ended scale, and category judgment gives subjects a small set of possible responses.

Some compelling evidence that response freedom may be more important than instructions comes from a small number of studies that have used category scaling instructions but allowed a wider range of responses than is usual. Rubin (cited in Marks, 1968, and S. S. Stevens & Galanter, 1957) allowed subjects up to 100 categories of response and found the category scale to become more like the magnitude estimation scale as the number of responses was increased. Marks (1968) found that increasing the number of categories tended to make the category scale more nearly like the magnitude scale. Gibson and Tomko (1972) compared category scales with either 7 or 49 steps and found the typical curvilinear relation between category and magnitude scales for the 7-category scale, but a linear relation with the 49-category scale. They concluded that it is the range of available responses rather than the instructions that determined the form of the category scale. Finally, Montgomery (1975) compared a number of methodological differences between magnitude and category scaling and found that range of response scale and freedom to choose the highest response number were the most important. Other factors, including the rule given the subject for responding, were negligible. When response freedom was allowed, subjects produced a magnitude estimation scale; when responses were restricted to a small set, subjects produced a category scale.

It is possible that the difference in response freedom in the two scaling techniques has not been connected often with the difference in results because there has been no available theory to explain why response freedom should make a difference. The present research supplies a theoretical foundation for the difference in the two subjective scales. If subjects perceive numerical magnitude in category scaling according to the function obtained in the present closed-range experiments, and if they perceive numbers in magnitude estimation according to the function obtained in the present experiments using an open-ended range, then the relation between magnitude estimation scales and category scales could be predicted solely from the different ways subjects use numbers in the

two cases (cf. also Baird, 1975; Banks & Hill, 1974). The prediction follows because of the nature of the numerical "rulers" subjects use under the two instructional sets. If the category scaling ruler is approximately linear (width of a unit does not change systematically with position on the scale) and the magnitude estimation ruler is compressive (width of a unit declines with positions higher in the scale), then the general curvilinear relation between the two scales is predicted. It does not matter what the rulers are measuring; if the category scale measurement for each stimulus is plotted as a function of the magnitude scale measurement, a curvilinear function will be obtained.

Figure 7 illustrates the sort of relationship between category and magnitude scales that would be predicted if subjects used subjective scales of number calibrated according to the closed-range function in category scaling and according to the open-range function in magnitude estimation. This figure plots the first through tenth order statistics for the 10 numbers generated in the 1-10 range of Experiment 6 (the results would have been virtually identical in form if we had used order statistics from either of the other ranges) plotted against 10 equally spaced points between 1 and 250 on the random-production subjective scale of Banks and Hill (1974, Experiment 2). The trend of the curve in Figure 7 is qualitatively the same as is found when category values for a given continuum are plotted as a function of magnitude estimation scale values for the same continuum. Variations in this general form would be expected if the category scale were distorted by stimulus spacing, end anchoring, and any of the other factors that can affect the form of the category scale. We therefore cannot make
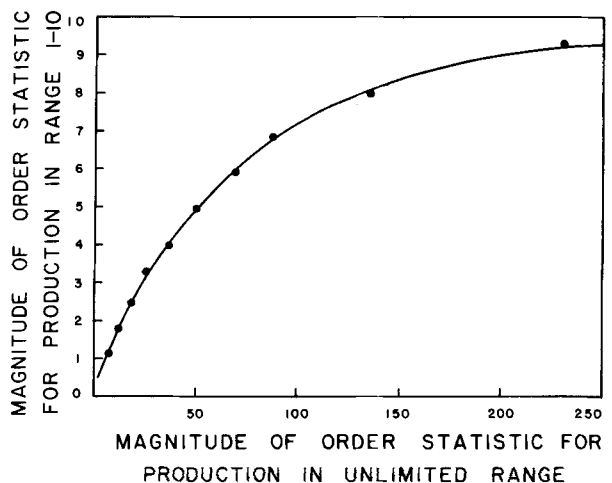

Figure 7. Relation between random production scales of number for productions given in an unlimited range (abscissa) and a limited range (ordinate). This relation is qualitatively the same as that between category judgment and magnitude estimation scales.

exact predictions about the form of the relation between category and magnitude scales without a model for the procedural factors that affect the form of the function. Nevertheless, Figure 7 gives a first-order approximation to the relationship.

The theory, then, is that subjects "measure" their sensations with a compressive ruler under magnitude estimation instructions and with an apparently linear ruler under category judgment. Does this mean that the category scale is more accurate than the magnitude scale? The answer to this question is negative, partly because of the biases that afflict category scales. Another reason is the uncertainty of generalizing from tasks in which numerical responses are unconstrained by stimuli to those in which they are. It is possible that requiring stimuli to be matched to the numbers will cause a systematic shift in the form of the subjective scale. The most that we can say, and the most that we need to say to predict the general form of the relation, is that the subjective number scale used in category scaling is less compressive than the subjective number scale used in magnitude estimation.

Aside from the relation to category scaling, what are the implications for magnitude estimation scales if the subjective scale of number is nonlinear? Banks and Hill (1974) have discussed some of these implications. If the number scale is a decelerated power function, as seems to be the case from the present research as well as Banks and Hill's, then power function scales derived from magnitude estimation will have exponents that are systematically inflated. If n is the exponent for subjective size of numbers (where $0 < n < 1.0$), then all magnitude estimation scales will have power function exponents larger by a factor of n^{-1} than the "true" subjective scale. Of course, it will be impossible to detect this from magnitude estimation scales, since their exponents will all be inflated by a common factor. Cross-modality matching, which is often considered to check the validity of magnitude estimation, would also be insensitive to a constant percentage change in magnitude estimation exponents (cf. Banks & Hill, 1974, pp. 372-375). In fact, as Banks and Hill note, a constant percentage change in magnitude estimation exponents would alter the conclusions of very few studies employing magnitude estimation or cross-modality matching. Most studies are concerned with the relative sizes of exponents or the goodness of fit of power functions, neither of which vary with a constant percentage change in the exponent.

The possible difference between the magnitude estimation function and a sensory scale uncontaminated by the numerical (or other) response scale becomes important when we consider a variety of techniques, such as equisection and bisection, that do not require a numerical response. Banks and Hill showed how

the results of these studies are consistent with the hypothesis that the "true" subjective scale is more compressive than the magnitude estimation scale, as would be expected if the numerical response scale multiplies the exponent by a factor greater than 1.0. Furthermore, the "input" scales found by Curtis and his colleagues (e.g., Curtis et al., 1968; Curtis & Fox, 1969) correspond to the "pure" sensory scale in this analysis, and these are also compressed relative to the magnitude estimation function.

We should also consider the scaling result that suggested the present studies (Marks & Slawson, 1966) because their findings seem to imply no difference between the magnitude estimation scale for loudness and the underlying sensory scale. The power function by which their tone bursts increased in intensity had an exponent of 1.67 for the series judged most linear. Such an exponent corresponds to a scale exponent of .6, which is the usual magnitude estimation value. However, their tones were spaced closely enough in time to be subject to effects of sensory interactions such as masking, summation, and loudness enhancement. If, for example, the loudness of each tone in the series was enhanced by the previous one (Zwislocki & Sokolich, 1974), then the effective subjective function by which successive tone bursts increased in loudness could have a larger power-function exponent than that measured in SPL by the experimenters. Thus, the series that increased according to an exponent of 1.6 could have had the same effect as one that rose with, say, an exponent of 2.0 but that was spread out enough in time to eliminate loudness enhancement (or other interactive effects, such as masking or summation). A rise function with an exponent higher than 1.6 implies a subjective scale more compressed than the magnitude estimation scale for loudness. Such a result is predicted by the hypothesis that magnitude estimation scales are expanded by the numerical response scale. To verify our interpretation of the Marks and Slawson results, it would be necessary to replicate their study over a number of interburst intervals.

One final question we should consider is the basis for the two scales. Why do we carry two scales around with us? One justification could be found in the fact that sometimes we need to be aware of percentage differences and sometimes we need to be aware of amount of difference. An easy way to approximate percentage differences is to measure with a compressive scale, whereas a linear scale is needed to record amount differences. It is possible that the two subjective scales developed in response to these two different cognitive needs. Why an open-ended scale tends to elicit the percentage metric (magnitude estimation) and why the restricted scale tends to elicit the amount metric (category judgment) are questions still to be answered.

REFERENCE NOTE

1. Borg, G., Edstrom, C-G., & Marklund, G. *A new method to determine the exponent for perceived force in physical work* (Reports from the Institute of Applied Psychology, No. 4). Stockholm: University of Stockholm, 1970.

REFERENCES

ATTNEAVE, F. Perception and related areas. In S. Koch (Ed.), *Psychology: A study of science* (Vol. 4). New York: McGraw-Hill, 1962.

BAIRD, J. C. Psychophysical study of numbers: V. Preferred state theory of matching functions. *Psychological Research*, 1975, **38**, 189-207.

BANKS, W. P., FUJII, M., & KAYRA-STUART, F. Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, **2**, 435-447.

BANKS, W. P., & HILL, D. K. The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology* (Monograph Series), 1974, **102**, 353-376.

CURTIS, D. W., ATTNEAVE, F., & HARRINGTON, T. L. A test of a two-stage model of magnitude judgment. *Perception & Psychophysics*, 1968, **3**, 25-31.

CURTIS, D. W., & FOX, B. E. Direct quantitative judgments of sums and a two-stage model for psychophysical judgments. *Perception & Psychophysics*, 1969, **5**, 89-93.

EISLER, H. Magnitude scales, category scales, and Fechnerian integration. *Psychological Review*, 1963, **70**, 243-253.

EKMAN, G. Is the power law a special case of Fechner's law? *Perceptual and Motor Skills*, 1964, **19**, 730.

EKMAN, G., & HOSMAN, B. Note on subjective scales of number. *Perceptual and Motor Skills*, 1965, **21**, 101-102.

GARNER, W. R., HAKE, H. W., & ERIKSEN, C. W. Operationism and the concept of perception. *Psychological Review*, 1956, **63**, 149-159.

GIBSON, R. H., & TOMKO, D. L. The relation between category and magnitude estimates of tactile intensity. *Perception & Psychophysics*, 1972, **12**, 135-138.

GOUDE, G. *On fundamental measurement in psychology*. Stockholm: Almqvist and Wiksell, 1962.

MARKS, L. E. Stimulus-range, number of categories, and form of the category scale. *American Journal of Psychology*, 1968, **81**, 467-479.

MARKS, L. E., & SLAWSON, A. W. Direct test of the power function for loudness. *Science*, 1966, **154**, 1036-1037.

MONTGOMERY, H. Direct estimation effect of methodological factors on scale type. *Scandinavian Journal of Psychology*, 1975, **10**, 19-29.

MOYER, R. S., & LANDAUER, T. K. Time required for judgments of numerical inequality. *Nature*, 1967, **215**, 1519-1520.

NOMA, E., & BAIRD, J. C. Psychophysical study of numbers: II. Theoretical models of number generation. *Psychological Research*, 1975, **38**, 81-95.

PETERSON, C. R., & BEACH, L. R. Man as an intuitive statistician. *Psychological Bulletin*, 1967, **68**, 29-46.

ROSNER, R. S. The power law and subjective scales of number. *Perceptual and Motor Skills*, 1965, **21**, 42.

RULE, S. J. Equal discriminability scale of number. *Journal of Experimental Psychology*, 1969, **79**, 35-38.

SCHNEIDER, B., PARKER, S., OSTROSKY, D., STEIN, D., & KANOW, G. A scale for the psychological magnitude of numbers. *Perception & Psychophysics*, 1974, **16**, 43-46.

STEVENS, J. C. Stimulus spacing and the judgment of loudness. *Journal of Experimental Psychology*, 1958, **56**, 146-150.

STEVENS, S. S. The direct estimation of sensory magnitudes—loudness. *American Journal of Psychology*, 1956, **69**, 1-25.

STEVENS, S. S. Issues in psychological measurement. *Psychological Review*, 1971, **78**, 426-450.

STEVENS, S. S., & GALANTER, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, **54**, 377-411.

WEISSMAN, S. M., HOLLINGSWORTH, S. R., & BAIRD, J. C. Psychophysical study of numbers: III. Methodological applications. *Psychological Research*, 1975, **38**, 97-115.

ZWISLOCKI, J. J., & SOKOLICH, W. G. On loudness enhancement of a tone burst by a preceding tone burst. *Perception & Psychophysics*, 1974, **16**, 87-90.