Journal of Experimental Psychology: General
1986, Vol. 115, No. 1, 62-75

0096-3445/86/$00.75

Can People Behave "Randomly?": The Role of Feedback

Allen Neuringer
Reed College

Experimental psychologists generally maintain that people cannot behave randomly. The present experiment asked students to generate random sequences of two numbers on the keyboard of a computer terminal. At first, all subjects' sequences differed significantly from random, thereby replicating the findings of the literature. But when given feedback from 5 or 10 statistical descriptors, the subjects learned to generate sequences that were indistinguishable, according to these statistics, from computer-generated random numbers. Randomlike behavior can therefore be learned.

When asked to behave randomly, people generally fail (for reviews, see Tune, 1964a, 1964b; Wagenaar, 1972). An experiment, by Bakan (1960) is a good illustration of the methods used in tests of "random" behavior. Seventy undergraduates were asked "to produce a series of 'heads' and 'tails' such as they might expect to occur if an unbiased coin were tossed in an unbiased manner for a total of 300 independent tosses." The subjects filled in "H" or "T" boxes on a form. Analysis of the frequencies of runs and triplets showed that responses differed from those expected by chance.

The failure of human subjects to behave randomly is a robust finding. The number of alternative responses have varied from 2, as in the work of Bakan (1960), through 3, 4, 5, 8, 10, 16, and 26 (Baddeley, 1966; Chapanis, 1953; Teraoka, 1963; Wagenaar, 1970a). Methods of responding have varied from calling out digits, letters of the alphabet, or nonsense syllables, to writing these same symbols on paper, pressing pushbuttons, touching metal disks with a stylus, or drawing lines on a paper (Baddeley, 1966; Lincoln & Alexander, 1955; Slak, Hirsch, & Syrja, 1979; Warren & Morin, 1965). Required speed of responding has been varied from self-pacing to from 0.25 s to 4 s per response (Baddeley, 1966; Teraoka, 1963; Warren & Morin, 1965). Age (Ross & Levy, 1958), mathematical sophistication of subjects (Chapanis, 1953), type of instructions (Beach & Swensson, 1967; Hyman & Jenkin, 1956), psychiatric evaluation of subjects (Horne, Evans, & Orne, 1982; Weiss, 1964), drug state (Truijens, Trumbo, & Wagenaar, 1976), and competing attentional demands (Evans, 1978; Evans & Graham, 1980) have been systematically varied. And a variety of statistics have been used as tests for randomness, with the most common being runs tests, chi-square, analyses of information content, and autocorrelation (Baddeley, 1966; Chapanis, 1953; Evans, 1978; Kuhl & Schonpflug, 1974; Lincoln & Alexander, 1955; Rath, 1966; Teraoka, 1963; Wagenaar, 1972). But no set of parameters reliably engenders random responding, despite clear instructions to subjects to behave as randomly as possible. Most researchers have therefore concluded that people do not behave randomly, and some have concluded that random behavior is impossible: "Producing a random series of responses is a difficult, if not impossible task to human [subjects], even when they are explicitly instructed . . ." (Wagenaar, 1971); "human [subjects] are incapable of generating a random series of selections from a finite number of alternatives . . ." (Tune, 1964a); "the human being is an extremely poor instrument for conduct of a random selection . . . Nor is this a quality that can be removed by conscious effort or training. Nearly every human being has as a part of his psychological makeup a tendency away from true randomness in his choice" (Yule & Kendall, 1950).

These conclusions are consistent with the determinism of much of contemporary psychology, from Freudian psychoanalysis to Skinnerian behaviorism. Stochastic processes are thought to be important (e.g., Estes, 1972), but the probabilistic nature of behavior is generally attributed to variability in environment or ignorance of experimenter, not to an inherent attribute of behavior. As knowledge is gained, we are told, precision of prediction will increase (e.g., Skinner, 1971).

It is clear that research performed over the last 50 years has not demonstrated human randomness. Whether or not people *can* behave randomly is less certain, for two different explanations are possible: explanation by trait and explanation by skill.

An explanation by trait implies that because of *inherent* limitations, people are incapable of random behavior. Such an explanation depends upon the fact that in a random series, responses must occur with approximately equal frequencies over the long run. So, too, all combinations of responses (of equal length) must be approximately equal. Different researchers posit different traits that limit human ability to behave randomly. According to one hypothesis, people fail because memory capacity is inadequate for retention of these frequencies (Baddeley, 1966; Tune, 1964a). Another hypothesis is that attentional processes do not permit subjects to ignore completely their previous responses, an inattention necessary, according to this view, for random behavior (e.g., Weiss, 1964). A third hypothesized limitation derives from subjects' difficulty in conceptualizing randomness: When presented with two series of numbers, subjects sometimes cannot discriminate random from nonrandom series (Wagenaar, 1970b; but see also Baddeley, 1966; Cook, 1967). The overwhelming agreement in the literature—that people do not generate random

I thank Rick Wood, Charles Green, and Gary Schlickeiser for invaluable technical assistance and Reed College staff and students for nurturing and supporting this research.

Correspondence concerning this article should be addressed to Allen Neuringer, Department of Psychology, Reed College, Portland, Oregon 97202.

sequences when requested—is taken as support for an explanation by trait, and most researchers have proceeded to search for responsible factors: What is it about human nature that makes people behave at least somewhat predictably?

On the other hand, explanation by skill has not received sufficient test. A hypothetical example will illustrate its relevance to randomness. Imagine an isolated society in which violins had never been seen or played, though violin music was often heard on the radio. As an experiment, a psychologist requested subjects to play a Beethoven violin concerto. Although the experimenter first asked the subjects if they knew the particular concerto, and all answered that they had heard the piece many times, when asked to play the piece, all failed. Imagine, moreover, that the research was replicated with parameters varied—for example, speed of playing, number of strings, age, knowledge, and drug states of the subjects, and that different measures of performance were employed, but failure to play was a robust finding. The conclusion, one analogous to the random generation case, was that people of the society are unable, perhaps because they lack the musical ability, requisite trait, or genetic precursor, to play the violin. However, for random performance, as for violin playing or any other complex skill, not *any* experience will suffice. Experience with dice, weatherman's estimates, or probability theory may no more suffice to teach the skill—if it is a skill—of random behavior than would studying a rule book enable a novice to play expert golf, or listening to the radio empower one to play a violin concerto.

There have been no direct tests of explanation by skill, but a few findings are suggestive. Chapanis (1953) reported that subjects sophisticated in mathematics generated numbers that were more nearly random than naive subjects. Ross and Levy (1958) found that young subjects behaved more nearly randomly than older subjects (indicating, perhaps, that people learn *not* to behave randomly), and that the older subjects (college students) became more nearly random after a class discussion concerning the nature of random ordering. The finding that people often do not discriminate random from nonrandom sequences suggests that changing the subjective definition of randomness may help people to behave randomly. A few operant conditioning studies also support an explanation by skill: Dolphins have been rewarded for novel jumps and flips (Pryor, Haag, & O'Reilly, 1969); rats for responding on two levers in a quasi-random fashion (Bryant & Church, 1974); pigeons for generating "least frequent" intervals between consecutive responses, the distribution of the birds' interresponse times eventually approximating a Poisson distribution (Blough, 1966); and pigeons for generating highly variable left–right patterns of responses (Page & Neuringer, 1985). These studies indicate that variability, and sometimes variability that meets criteria of "randomness," can be reinforced in animals.

The present study tested the widely accepted conclusion that people are *unable* to behave randomly by evaluating whether feedback would enable subjects to learn to generate random sequences. Students sat at a computer terminal, generated sequences of "1s" and "2s," and received statistical feedback after each set of 100 responses. This attempt to reinforce human randomness was a direct test of the skill theory of random performance.

A brief word must be said about the meaning of "randomness," a concept for which, unfortunately, there is no easy definition (see, for example, Feller, 1968; Popper, 1968; von Mises, 1957). In its simplest sense, randomness implies (a) equal probability of alternative events or combinations of events and (b) the inability of an observer to improve the prediction of the next event from knowledge of any previous set of events. A problem arises, however, when one attempts to decide whether a particular finite sequence can be described as random or not. Mathematical discussions of randomness generally refer to infinite sequences. In an infinite random sequence, any particular finite sequence is possible. Indeed, every finite sequence is exactly as likely as every other sequence of same length. For example, in an infinite sequence of 1s and 2s, it is possible to find a sub-sequence that consists of one hundred 1s in a row, or, indeed, one million 1s, and the sub-sequence consisting of one million 1s is exactly as likely as any other particular sub-sequence of one million digits—for example, 1112122112122 . . . (see Lopes, 1982). These considerations indicate the impossibility of proving with certainty that a particular finite sequence *deviates* from random—that is, that the finite sequence was not selected from an infinite random series. The second side to this coin is that no matter how many statistical evaluations indicate that a finite sequence *is* random, there may exist some other test that shows nonrandomness. There is no conclusive test, or set of tests, to prove the randomness of a finite sequence: The null hypothesis cannot be proven (see Chaitin, 1975, for a related argument).

Experimental psychologists have dealt with this problem by using common statistical methods to evaluate the probability that a particular sequence had been selected from a random population. For example, if 1,000 samples were independently selected from an infinite random sequence of 1s and 2s, with each sample containing 100 numbers, many more of the 1,000 samples would contain approximately 25% of each of the four pairs, 1-1 (i.e., 1 followed by 1), 1-2, 2-1, and 2-2, than would contain all 1s and none of the other three pairs. Statistical tests of particular attributes of a finite sequence enable statements concerning the probability that the sequence was selected from a random population.

In the present research, performance of a human subject was called random if it was statistically indistinguishable from that of a simulating computer-based random number generator under analogous conditions. In most previous research on human randomness, few statistics, often one or two, were used to show that the subjects' responses were not random. The present strategy was also to choose a small set of statistical tests that showed that subjects initially deviated from random. The question then was whether, through training, performances could be modified so that the person became statistically indistinguishable from the random generator on the chosen tests. Human randomness was therefore defined by a variant of the Turing game. If the human's performance could not be distinguished from a random generator by common statistical analyses, the human was described as behaving randomly.

Experiment 1

Method

Subjects

Seven Reed College undergraduate students served—five females (D, H, P, R, and S) and two males (Sh, and Y).

Table 1
Representation of RNG1 Responses

	Response $n+1$		
Response n	1	2	Sum
1	3	4	7
2	3	2	5

Apparatus

An Osborne 1 computer, containing an alphanumeric keypad and attached to a 12-in. (30.5-cm) Zenith Data Systems monitor, was located on a desk in a laboratory room.

Procedure

Baseline condition. We first sought to establish whether, as in almost all previous studies, responses were *not* random in the absence of feedback. Subjects were told to press the "1" and "2" keys on the alphanumeric keypad as randomly as possible. Pressing any other key produced a brief error message. There were no constraints on response speed. A single session, lasting approximately 1 hour, provided 60 trials of 100 responses each for a total of 6,000 baseline responses. Each 100-response trial was terminated with the screen message "TRIAL OVER; Next Trial is ##; Please Continue," with the "##" containing the trial number and accompanied by a double beep. There was no other feedback. Subjects were paid $3.50 for participating.

Feedback condition. The conditions were the same as baseline except that feedback from five statistical descriptors was given after each trial. The subjects were asked to vary their responses so that distributions of these five statistics would approximate those calculated from a random number generator under analogous simulated conditions. The five statistics were as follows.

1. RNG1: This descriptor, evaluating the amount of information in a sequence of responses (Evans, 1978; Miller & Frick, 1949; Tulving, 1962), was based on the frequencies of pairs of contiguous responses, namely 1-1 (i.e., 1 followed by 1), 1-2, 2-1, and 2-2. The closer to equal these four pairs in a given trial, the closer RNG1 would be to 0.0; the more unequal the frequencies of the four pairs, the closer RNG1 would be to 1.0. For example, the frequencies of consecutive pairs in the sequence—1, 2, 2, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2—can be represented as in Table 1. The marginal numbers show the approximate sum of the 1s (7) and 2s (5). (Note that the marginal frequencies will always be less by one than the total of 1s and 2s due to the fact that the last response is not followed by any other. This is not serious when the total number of responses is large, e.g., 100.) The RNG1 index was computed as follows: RNG1 = $\Sigma C^*\log(C) - K/\Sigma M^*\log(M - K$, where C refers to the frequencies in each of the 4 cells and M refers to the marginal frequencies. K is a constant equal to $\Sigma C^*\log(C)$ when frequencies in each cell are as equal as possible (K = 137.97 in the present case).

2. RNG2: This descriptor was identical to RNG1 except that instead of contiguous responses, every other response was used to define response "pairs." Thus, in the sequence 1, 1, 2, 1, 1, 1 . . . , the response pairs entered into the RNG2 table were, in order, 1-2, 1-1, 2-1, 1-1 and so forth. Because there were only 98 pairs per 100 responses, K equaled 136.15.

3. Alternations (ALTS): The third statistic described the number of runs, defined as a sequence of 1s followed by a 2, or a sequence of 2s followed by a 1. A sequence includes a single instance. Thus, in the following set of numbers, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, there are a total of 7 runs, or ALTS. The maximum possible number of ALTS was 99 (where 1 and 2 alternated throughout a session), and the minimum was 0.

4. C1: This descriptor, analogous to RNG, compared performances on Trial n with that on Trial $n - 1$, so as to test whether a single sequence or strategy was being learned. The pairs of numbers entered into the table were generated from analogous response positions across two consecutive trials. Thus, the first response in Trial n and the first response in Trial $n - 1$ generated the first pair, the second response in Trial n and second response in Trial $n - 1$ constituted Pair 2, and so on. This descriptor therefore evaluated the consistency, or randomness, of responses *across* trials.

5. C2: This descriptor was identical to C1 except that the pairs of responses were generated from a comparison of every other trial. Thus, the first pair of responses was derived from the first response in Trial n and first response in Trial $n - 2$, and so on through the 100 responses.

Feedback table. To provide subjects with feedback relating their performance to the random generator, a feedback table was constructed in the following way. First, using the random generating procedure internal to the Osborne computer, 20,000 1s and 2s were generated—2,000 trials consisting of 100 numbers each. The data in each trial were analyzed according to the above five descriptors, thereby yielding a total of 2,000 values for each of the descriptors. These 2,000 values per descriptor were then ordered (separately for each descriptor) from highest to lowest, and 20th-percentile boundaries were calculated. Thus, for each descriptor there were four boundaries, defining five equal classes. These 20th-percentile boundaries were used to establish the five classes (CL1 through CL5) shown across the top of Table 2.

The subjects were asked to learn to generate sequences of 1s and 2s yielding descriptor values that fell approximately equally across the five classes associated with each of the five descriptors. By so doing, the subjects would be approximating the statistical distributions of the random generator. To take RNG1 as example, if a subject's performance on a given trial generated a value for RNG1 that was lower than the first boundary calculated from the random number generator, then the figure in the cell at the intersection of the CL1 (read "Class 1") column and RNG1 row would be increased by 1. If the subject's performance generated an RNG1 value that was between the first and second boundary points, there would be an increase of 1 in the CL2 column. The numbers in the cells of the table were cumulated across a given session, with each session containing 60 trials. Finally, so that subjects could know how they had performed on the last trial, the increment on the last trial was underlined. For example, if a subject's last trial RNG1 score fell in the CL1 category, then the CL1 score was underlined. Table 2 provides an example of the feedback seen by a subject after 32 trials. On the last trial, Class 1 was incremented in the RNG1 row, Class 4 in the RNG2 row, and so forth. Note that the C1 descriptor value could not be calculated on the first trial, and therefore the sum of C1 across the five classes was always one less than the three descriptors lying above it, and, similarly, C2 contained two fewer entries.

Table 2
An Example of the Feedback Table Presented to Subjects Following Each Trial in Experiment 1

Descriptor	CL1	CL2	CL3	CL4	CL5
RNG1 (var)	9	6	6	4	7
RNG2 (var)	10	3	6	5	8
ALTS (few)	8	11	5	3	5
C1 (low C)	7	6	8	3	7
C2 (low C)	8	7	6	2	7

Note. CL = Class. Parenthetical comments indicate Class 1 performance (e.g., in the case of ALTS, very *few* alternations would result in an increment of CL1). See text for description of each of the five statistical descriptors (RNG1, RNG2, ALTS, C1, C2).

In brief, after each trial, one cell was incremented along each of the five rows (statistical descriptors); when a particular cell was incremented, that cell was underlined. Subjects attempted to equalize cell values across each of the five rows so as to approximate a random generator.

In the session following the no-feedback baseline condition, feedback was provided for only the RNG1 descriptor. Once a subject's performance generated instances in each of the five classes of RNG1, the next descriptor, or RNG2, was added, and so forth until feedback on all five descriptors was provided for the remaining trials. As each descriptor was added, its function was briefly described. The feedback table remained in view on the monitor, although unchanging, throughout each trial. At the end of the trial, as in the baseline condition, the screen blanked out, "TRIAL OVER" appeared, followed by the updated table, and "Next Trial is ##. Please Continue." Again, as in baseline, subjects were free to respond at any speed and to spend as much time as they chose examining the feedback table between trials. Sessions generally terminated after 60 trials.

All questions were answered truthfully as to how the various descriptors were calculated. Furthermore, the experimenter continually suggested ways to improve performance. These suggestions were both specific (e.g., "The ALTS data show that you are not repeating one or another response often enough, but are jumping back and forth too often") and general (e.g., "Imagine a spring that is pulled to the left whenever you emit a '1' and to the right whenever you emit a '2.' Over the long run, the spring wants to be at rest, but to be random, you must build up intermittent tension"). The one exception to this verbal feedback was during the session in which the subject's performance was defined as "random." During that session, the experimenter was either absent from the room or sat quietly, offering no advice and responding to no questions.

At the beginning of the feedback procedure, subjects were told that they would be paid $2.50 per hour for their participation and that if they learned to "be random" they would receive an additional $15.00. All money was disbursed at the completion of the experiment. Sessions continued until a subject was evaluated as "random" over 60 consecutive trials according to all five statistics.

Statistical test for randomness. Kolmogorov-Smirnov (K-S) tests (Siegel, 1956) were used to compare a given subject's data across 60 trials to the random number generator. (Because we were attempting to test whether individual subjects could learn to behave randomly, averaging across subjects would have been misleading and inappropriate.) A separate K-S test evaluated each of the five descriptors. The random number generator was programmed to generate 1,000 new trials, a number large enough to approximate the theoretical distributions; for each of these 1,000 trials the five descriptors were calculated, and these were then employed as the comparison data in the K-S tests. If, for a given descriptor, the human subject's data could not be statistically distinguished from the random generator at the $p = .05$ level of significance, the data were said to be random according to that descriptive statistic. The K-S test was chosen because it makes no assumptions about the distribution of scores—that is, it is a nonparametric statistic, but it evaluates differences in distributions as well as in central tendencies. When there were no significant differences between subject and random generator on any of the five descriptors, the subject was said to have successfully generated numbers randomly and consequently received the additional $15.00 plus whatever money accumulated over the sessions.

Results

The first question was whether performances differed statistically from random during baseline, where no feedback was provided. Baseline scores on each of the five descriptive statistics were compared by the K-S tests with the analogous scores from the computer-based random generator. All subjects differed significantly ($p < .05$) from the random generator, with two subjects differing significantly on all five descriptors, two subjects differing on four descriptors, two subjects on three descriptors, and the last subject on two descriptors. In terms of the particular descriptors involved, all seven subjects differed significantly on both RNG1 and RNG2, and four of the subjects differed significantly on each of the remaining three descriptors. Since "random" implies that responses do not differ from chance on *any* of the statistics, the first conclusion is that all subjects were "not random" during baseline.

Subjects were then given feedback on the five descriptive statistics, in the order shown in Table 2. The number of trials before each subject received feedback on the complete set of five descriptive statistics were, in alphabetical order of name, 131, 81, 156, 91, 130, 91, and 215, respectively.

The second main question was whether subjects could learn to produce sequences that did not differ significantly, according to the five descriptors, from the random generator. During the last 60 trials of the feedback condition, all subjects were statistically indistinguishable ($p > .05$) from the random generator on all five K-S tests. The number of trials since the beginning of feedback before the random criterion was met was, again in alphabetical order of names, 216, 261, 171, 166, 483, 186, and 385 for the seven subjects, respectively.

Figure 1 shows one example of performance changes across trials. Depicted are all subjects' RNG1 scores divided by the average RNG1 score from the random number generator. A value of 1 indicates that the subject's score equaled the random generator's score, a value of 2 indicates that the subject's score was twice that of the random generator, and so on. Note that the ordinate is logarithmic, so that ratios greater and less than 1 would be similarly represented. Each point is a median of a block of 20 trials, with trials along the abscissa. In all cases, RNG1 scores were initially high relative to the average of the random number generator's RNG1—that is, scores were greater than 1.0. Furthermore, there was a tendency during baseline for RNG1 scores to deviate *increasingly* from the random generator. However, with feedback, RNG1 scores decreased and approached equality with the random generator. All subjects behaved in approximately the same manner.

Figure 2 shows values from one subject of each of the five statistics across all sessions. These data are generally representative of all subjects. Again, for each descriptor, medians of blocks of 20 trials are divided by the average random number generator's score on the given descriptor. Thus, 1.0 along the ordinate, through which the dashed line is drawn, again indicates that the subject's performance equaled the average of the random number generator's performance. RNG1 and RNG2 both started above the random number generator and decreased across sessions. Similarly, C1 and C2 began too high. In these four cases, therefore, the subject's responses, both within trials and across trials, were considerably more patterned or repetitive than was the random generator. ALTS, which indicate the number of times the subject alternated between a series of 1s and a series of 2s, started at lower than random, indicating too few alternations, but over trials, ALTS increased systematically. Most previous research in this area (see Wagenaar, 1972) showed that naive subjects emit *too many* alternations. The present findings, as well as a few others in the literature (e.g., Weiss, 1964) may be due to the type of response required. It was easy to respond very rapidly on a single response key. In many of the previous experiments, the

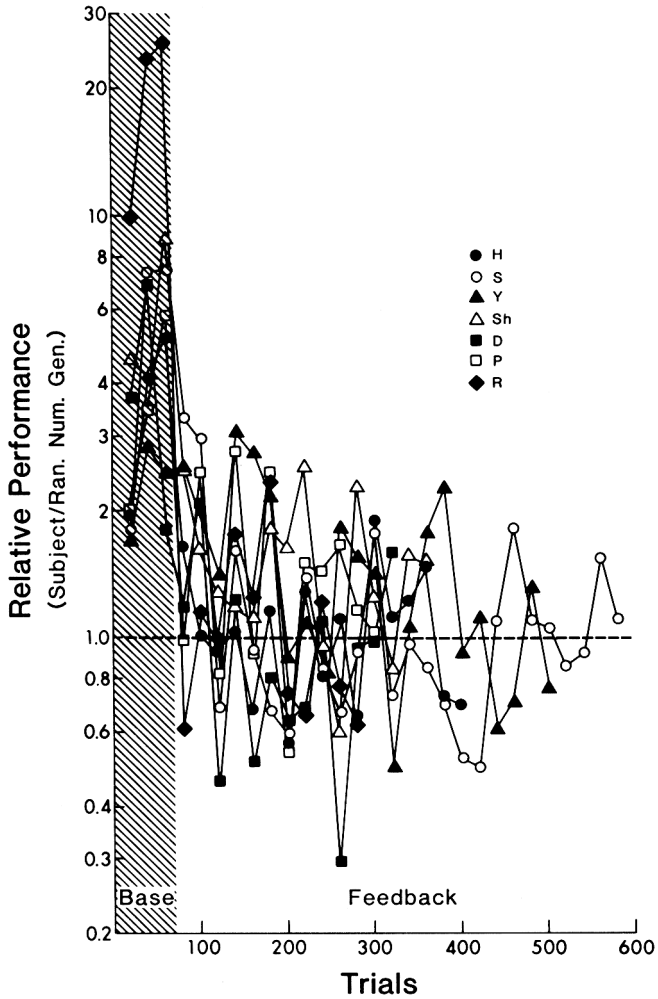


Figure 1. RNG1 scores for each subject relative to the random number generator's average score. (Each point represents the average of a block of 20 trials.)

response topography may have combined with a required interresponse time to increase the likelihood of alternations.

The main result was that after less than an average of 6 hours of feedback training, seven subjects who had been statistically nonrandom now behaved randomly according to the five statistical comparisons employed.

Experiment 2

Experiment 2 attempted to replicate Experiment 1 except with 10 statistical descriptors, rather than 5, and with a more demanding statistical evaluation of "random." Also, advice and guidance given by the experimenter were minimized.

Method

Subjects

Four high-school students, S (a male), F, W, and Y (3 females), spent approximately 1 hour per day, 5 days per week, in the present experiment. The four students were paid participants in a summer science program for minority students, and this experiment was part of their duties. The subjects had no previous statistical training.

Apparatus

The subjects sat at 4 Digital Decscope terminals in a room containing 11 such terminals connected to a PDP-1170 computer. The terminals had alphanumeric keypads, on which responses were entered.

Procedure

The experimenter generally sat in the room, but experimenter interactions with subjects were confined to procedural questions and to restarting sessions after a breakdown or upon request of the subjects.

Baseline. The subjects were instructed to "behave as if you were a tossed coin" and enter the digits "1" or "0" as randomly as possible. As in Experiment 1, each subject then generated 6,000 1s or 0s, divided into 60 trials of 100 responses each. At the end of each trial, "Trial Over, Please Hit Return to Continue" appeared on the screen and, when the return key was pressed, there appeared "Please Begin." No other feedback was given except that if any key other than "1" or "0" was pressed, a "beep" sounded, and "Please Type 0 or 1" appeared on the screen. Subjects were free to respond at any speed. The baseline session took approximately 1 hour.

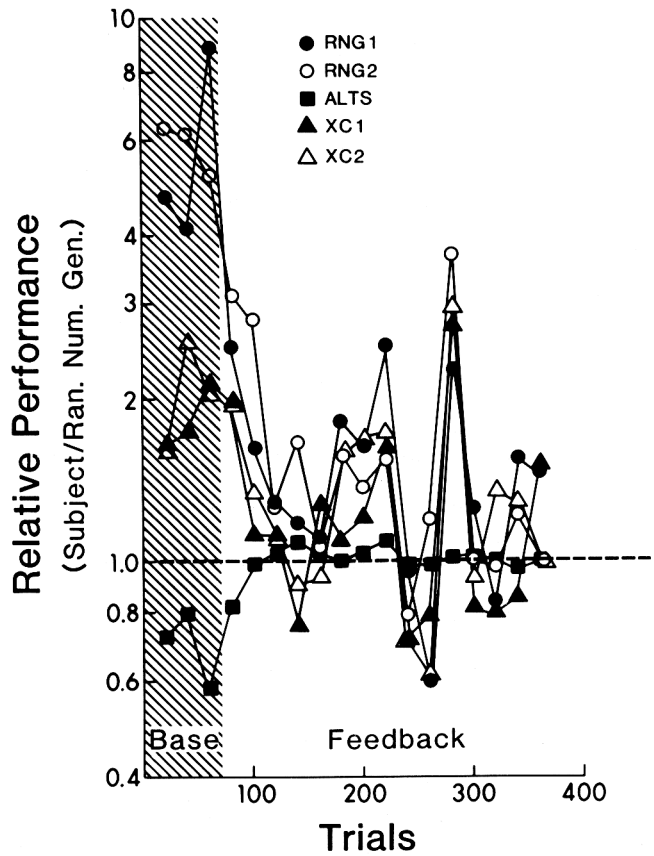


Figure 2. Performance by subject Sh as evaluated by each of the 5 descriptive statistics relative to the average of the random number generator's scores. Each point is an average over 20 trials.

Feedback #1 (Kolmogorov-Smirnov). There were 20–25 sessions under this condition, each session lasting approximately 1 hour. As in baseline, subjects generated sets of 100 responses (0s or 1s) per trial, at a self-paced speed. Unlike baseline, however, feedback in the form of a table of numbers appeared on the screen at the end of each trial. Table 3 shows an example of the feedback table, with each line representing a different descriptive statistic. This feedback was analogous to that in Experiment 1, but there were now 10 statistical descriptors. To determine the 20th-percentile boundary points, we programmed a random number generator (a standard linear congruential generator, as described by Knuth, 1969) to generate 100,000 responses (0s or 1s). Each set of 100 responses constituted one trial, and every trial yielded a score for each of the 10 statistical descriptors. The resulting 1,000 values were then ordered from highest to lowest to create five equal classes for each descriptor (shown in Table 3). At the end of each trial, the subject received feedback analogous to that in Experiment 1. After reviewing the feedback table, the subject pressed the return key and another trial was initiated, with "Please Begin" appearing on the screen.

Before the first feedback session, the subjects were told that their task was to learn to generate numbers randomly such that the five classes for each of the descriptors would contain approximately equal instances. Over the course of the first five sessions following baseline, the feedback given to the subjects was increased from 1 to 10 descriptors. As each descriptor was added, the experimenter briefly described how it was calculated and how different performances might affect it. By the 450th trial, all subjects were receiving feedback on all 10 descriptors. Thereafter, throughout the remainder of the experiment, each trial was followed by feedback on all 10 descriptors. At this point, the subjects were told that they would receive 2 days off with pay if they attained a level of performance on all 10 descriptors such that a K-S comparison showed them to be "random" over two successive sets of 60 trials each. The K-S test was the same as in Experiment 1, but the 1,000 comparison "theoretical" scores were regenerated. At the end of each session, the experimenter briefly pointed out those descriptors for which there were significant differences between subject and random number generator. This condition continued for the four subjects, in alphabetic order of their initials, for 927, 909, 1,772, and 1,200 trials, respectively.

The 10 descriptors, in the order of their appearance in the feedback table, were as follows:

Percent 0 (%0) indicated the percentage of 0s (number of 0s divided by total responses) emitted in a trial. Percent 0s could vary from 0% to 100%.

ALTS represented the number of times per session that a 0 was followed by a 1, or vice-versa.

Lines 3 through 6 represented runs. RUNS1 are cases in which a single 1 is followed by one or more 0s, or a single 0 followed by one or more 1s. In the following sequence there are five runs of Length 1: 0010111011010000001. In the same sequence, there are two instances of RUNS2, one instance of RUNS3, and one instance of RUNS4–10, the latter including runs of length 4 through length 10. The possible range of RUNS1 was 0 through 99, of RUNS2 was 0 through 49, of RUNS3 was 0 through 33 and of RUNS4–10 was 0 through 24.

The next two lines contained RNG1 and RNG2, these being identical to the RNG measures used in Experiment 1.

The final two lines presented C1 and C2, these again being identical to C1 and C2 in Experiment 1.

Feedback #2 (Kolmogorov-Smirnov and t test comparisons). In Experiment 1 and in the Feedback #1 condition above, CL1 and CL5, the lowest and highest categories, had no lower and upper bounds, respectively. Thus, for example, if a subject had tended to generate too many runs of Length 1, thereby skewing the feedback table toward CL5, the number in the CL1 category could be increased by omitting all runs of Length 1. Although such a strategy would not adversely affect the Kolmogorov-Smirnov nonparametric comparison, the generated values could lie outside the range of values produced by the random number generator. Analysis of the data showed that although subjects were statistically indistinguishable from the random generator under the K-S tests, they were in fact producing values that lay outside the range of the random generator's values. Therefore, the number of classes in the feedback table was increased from five to seven, with the lowest and highest classes being defined as "no low" and "no high," respectively. To accomplish this, two additional columns were added to the table, a column to the left of CL1 labeled "no low" and a column to the right of CL5 labeled "no high." In all other respects, the table was identical to that in Table 3. New categories were defined as follows: The 1,000 trials generated by the random number generator were ordered from lowest to highest on each of the 10 descriptors separately. Now, however, rather than dividing the 1,000 numbers into 5 twentieth-percentile classes (200 per class), the lowest category was obtained by counting up 20 cases from the lowest value, and the highest category was obtained by counting down 20 cases from the highest value of the random generator's values. Thus, the lowest category contained the lowest 2% and the highest category the highest 2%. The remaining five categories each contained 19.2% of the random generator descriptor scores (as opposed to the 20% in Table 3). Subjects were asked to equalize their performances across the middle five categories but to get few if any instances in the "no low" and "no high" categories. (These categories were referred to as *no* low and high, respectively, because pilot work showed that it was difficult for subjects to avoid these categories. In fact, 2% of instances were acceptable within each of these extreme categories.) As a second major change in procedure, after each set of 60 trials, the subjects were told whether or not they differed significantly from the random generator according to the previously used Kolmogorov-Smirnov test, as well as according to a t test, which takes into account the absolute values of the outlying data. Although some of the present data do not meet the requirements for a t test statistic (e.g., normal distribution), it was found through experimental iteration that a t test was more demanding than the K-S test (i.e., there were many instances in the Feedback #1 condition in which the K-S test showed no statistical difference between subject and random generator but where the t test indicated a statistically significant difference). The subject's task was to generate 1s and 0s over 60 consecutive trials that did not differ from the random generator at $p = .05$ level under both Kolmogorov-Smirnov and t test evaluations. Sessions were continued until the subject met the contingency.

Table 3

An Example of the Feedback Table Given to Subjects Following Each Trial in Experiment 2

Descriptive statistic	CL1	CL2	CL3	CL4	CL5
%0 (low)	11	14	8	12	13
CHANGE (low)	9	11	10	18	8
RUN1 (low)	13	13	9	12	12
RUN2 (low)	10	10	18	8	11
RUN3 (low)	12	12	12	12	12
RUN4–10 (low)	15	9	11	8	15
RNG1 (var)	13	13	6	14	12
RNG2 (var)	11	11	10	12	13
C1 (low)	7	12	18	9	11
C2 (low)	6	17	11	11	10

Note. CL = Class. Parenthetical comments indicate class 1 performance (e.g., in the case of %0, very few zero responses would result in an increment of CL1). See text for description of each of the 10 statistical descriptors.

Results

The main results were that all subjects differed significantly from the random generator during baseline, and all then learned to behave randomly as assessed by both K-S and *t* tests.

K-S tests compared the random number generator (1,000 scores) to the subjects' scores during the 60 trials of the baseline condition. For three of the four subjects, all 10 statistical descriptors differed significantly from the random number generator at the $p = .05$ level. For Subject F, 8 of the 10 descriptors differed significantly, with %0 and C1 not significant (see Table 4). Thus, all subjects generated numbers that were clearly "not random." This again replicated the robust finding that people do not behave randomly upon request.

After receiving feedback in the Feedback #1 condition, the four subjects met the requirement of no statistical difference (K-S tests) between subject and random generator on all 10 descriptors over two successive sets of 60 trials. Subjects F, S, W, and Y met these requirements after 926, 848, 1,771, and 1,020 trials, respectively.

However, each of the subjects differed on one or more of the 10 descriptors when *t* tests were used to compare subject with random generator. A review of the subjects' data indicated that a major difference between their behavior and the random generator was the many extreme or outlying values in the subjects' data. Therefore, the "no low" and "no high" categories were added, with the result that performances became statistically indistinguishable from the random generator along all 10 descriptors using both K-S and *t* test evaluations (see Table 4 for exact *t* and *p* values). Subjects F, S, W, and Y attained 60 consecutive trials of random performance (under both K-S and *t* test) after an additional 796, 1,143, 203, and 240 trials, respectively. Providing feedback regarding outlying values helped subjects to approximate more precisely the random generator.

Table 4
t and p Values for t tests and p Values for Kolmogorov-Smirnov Tests Under Baseline (B) and Final 60 Sessions (R)

	F			S			W			Y		
	t test		K-S test	*t* test		K-S test	*t* test		K-S test	*t* test		K-S test
Descriptor	*t*	*p*	*p*	*t*	*p*	*p*	*t*	*p*	*p*	*t*	*p*	*p*
%0												
B	**.111**	**.908**	**.10**	**1.768**	**.074**	.025	**1.563**	**.114**	.001	5.657	.000	.001
R	**1.117**	**.263**	**.10**	**.873**	**.613**	**.10**	**.850**	**.600**	**.10**	**.999**	**.681**	**.10**
ALT												
B	7.264	.000	.001	8.444	.000	.001	18.829	.000	.001	13.863	.000	.001
R	**.758**	**.545**	**.10**	**.268**	**.785**	**.10**	**.556**	**.585**	**.10**	**.222**	**.819**	**.10**
RUN 1												
B	3.524	.029	.025	**1.115**	**.264**	.001	8.484	.000	.001	12.461	.000	.001
R	**1.298**	**.191**	**.10**	**.119**	**.901**	**.10**	**.541**	**.595**	**.10**	**.870**	**.612**	**.10**
RUN 2												
B	18.625	.000	.001	9.079	.000	.001	13.056	.000	.001	8.539	.000	.001
R	**.399**	**.693**	**.10**	**.699**	**.508**	**.10**	**1.267**	**.203**	**.10**	**1.646**	**.096**	**.10**
RUN 3												
B	9.329	.000	.001	8.581	.000	.001	10.226	.000	.001	4.972	.000	.005
R	**.685**	**.501**	**.10**	**.408**	**.687**	**.10**	**.593**	**.561**	**.10**	**.686**	**.500**	**.10**
RUN 4												
B	11.245	.000	.001	2.351	.018	.01	**1.771**	**.073**	.001	9.184	.000	.001
R	**.451**	**.657**	**.10**	**.363**	**.718**	**.10**	**.298**	**.763**	**.10**	**.434**	**.668**	**.10**
RNG 1												
B	8.239	.000	.001	20.232	.000	.001	43.147	.000	.001	13.454	.000	.001
R	**.638**	**.531**	**.10**	**.664**	**.514**	**.10**	**.181**	**.851**	**.10**	**1.894**	**.055**	**.10**
RNG 2												
B	24.607	.000	.001	18.138	.000	.001	47.014	.000	.001	3.958	.002	.005
R	**.634**	**.534**	**.10**	**.578**	**.576**	**.10**	**.015**	**.985**	**.10**	**.241**	**.805**	**.10**
C1												
B	**1.428**	**.150**	**.10**	8.882	.000	.001	30.231	.000	.001	2.829	.005	.025
R	**.107**	**.912**	**.10**	**.587**	**.565**	**.10**	**.604**	**.553**	**.10**	**.420**	**.678**	**.10**
C2												
B	2.776	.006	.05	7.604	.000	.001	31.714	.000	.001	**1.539**	**.120**	.005
R	**.699**	**.509**	**.10**	**.631**	**.536**	**.10**	**1.080**	**.280**	**.10**	**.597**	**.558**	**.10**

Note. F, S, W, and Y are the initials of the four subjects. Each test is based on 60 values from a human subject and 1000 values from the random number generator. Bold figures indicate no statistical difference between people and random generator. The .10 value under K-S indicates $p > .10$. See text for description of each of the 10 statistical descriptors.

Figures 3 through 6 show how closely the distributions of the subjects' descriptor values approximated the random values. In each figure, the left column, marked Base, shows performances during the 60 baseline trials, and the right column, marked RAND, shows performances during the 60 trials when the subject was evaluated as random according to both K-S and *t* tests under the Feedback #2 conditions. Each of the curves is a frequency distribution of the descriptive statistic scores, the solid lines (Xs) representing a subject's performance and the dotted lines (Os) showing the random number generator. Along each abscissa are 11 categories, generated by dividing the range obtained from the random number generator's 1,000 scores by 10 and using the resulting interval as category size. The ordinates show the frequencies of instances in each of the 11 categories. Note that the frequencies for the random number generator were obtained from 1,000 trials, so as to approximate theoretical distributions, and then scaled down to equal the 60 instances of the subjects. Note, also, that each ordinate was scaled to show maximum detail.

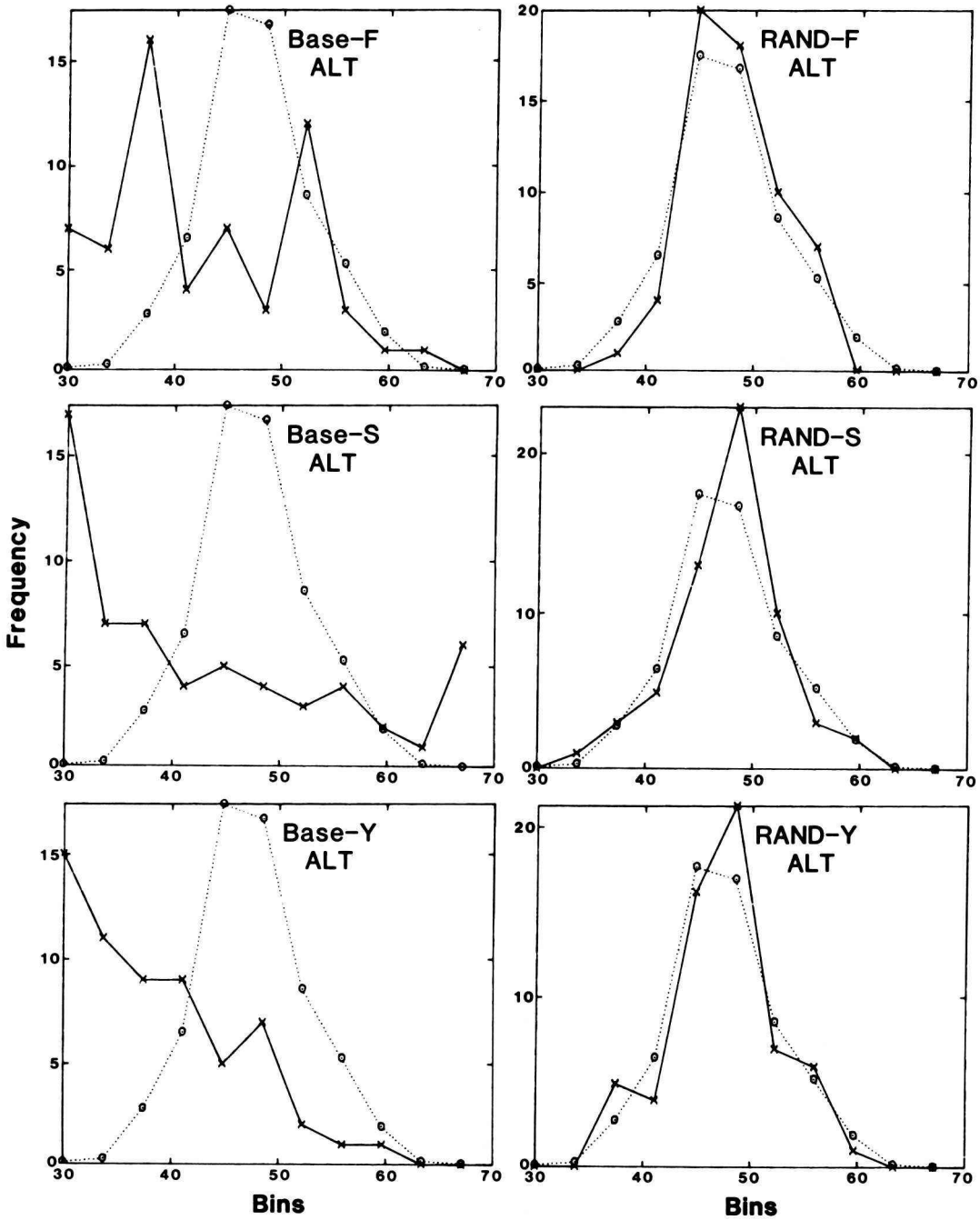


Figure 3. Frequency distributions of ALT scores for 3 subjects (F, S, and Y) in Baseline (Base column) and after learning to behave randomly (Rand column). (For comparison, the dotted line shows performance of the random number generator.)

Therefore, for example, the curves for the random generator in all 6 boxes of Figure 3 are identical, with only the ordinate scale differing. Comparisons between subject's data and random generator can easily be made directly on any given graph. Of main interest is whether and how the distributions of the subjects' descriptor scores changed with training.

Figure 3 shows the ALT descriptor for three subjects, F at top, S in the middle row, and Y at bottom. The middle row of Figure 4 shows these same data for subject W. The left-hand columns show that the frequency distributions of the ALT descriptor scores differed greatly from that of the random generator during baseline: The random generator's scores were normally distributed, whereas the subjects' distributions clearly were not. The right-hand columns show that during the final 60 trials, the subjects' distributions closely approximated the normal distribution of the random generator.

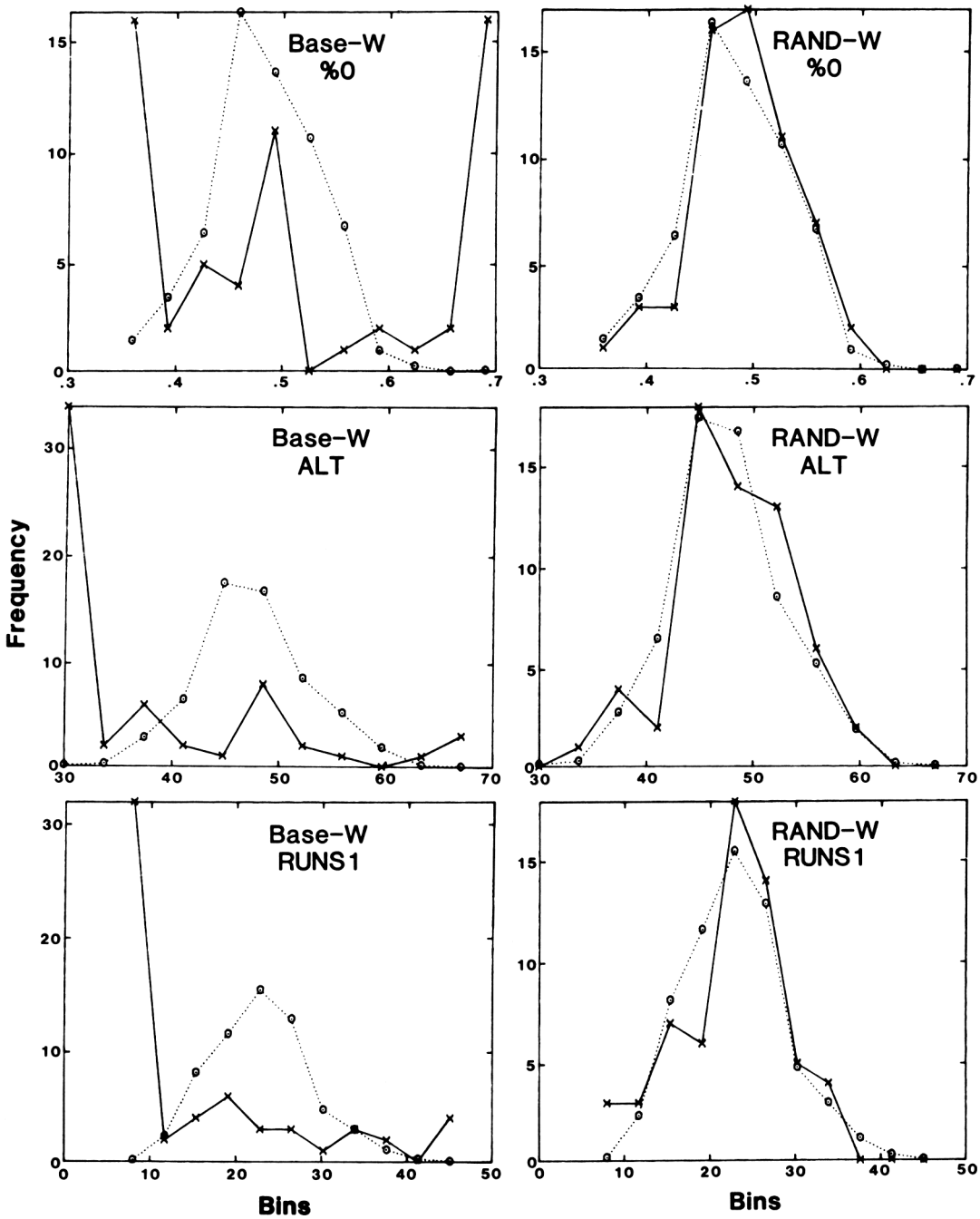


Figure 4. Frequency distributions for Subject W on three descriptive statistics: percent zero, alternations, and runs of length 1. (Left column shows performances under baseline condition and the right column after subject had learned to behave randomly. The dotted line shows comparable performance by the random number generator.)

To conserve space, Figures 4, 5, and 6 represent the distributions of 9 descriptors for one subject, W. In each case, the subject's distributions during baseline (left-hand column) differed markedly from the random generator. Note, especially, the RNG1, RNG2, and C1 data shown in Figure 6 (C2 was omitted to conserve space but is essentially the same as C1), where the subject's distributions were *opposite* those of the random number generator during baseline. In these cases, the random number generator's distributions were exponentially *decreasing* in form, but the human subjects began the experiment with these distributions *increasing* in form. By the end of the experiment, the distributions of all 10 statistical descriptors closely approximated the distributions from the random generator (right-hand column). In summary, initial distributions of subjects' descriptor scores differed both from the random number generator and often from one another. By the end of the experiment, the distributions were practically identical, both for subject compared with subject and for subject compared with random generator. According to the

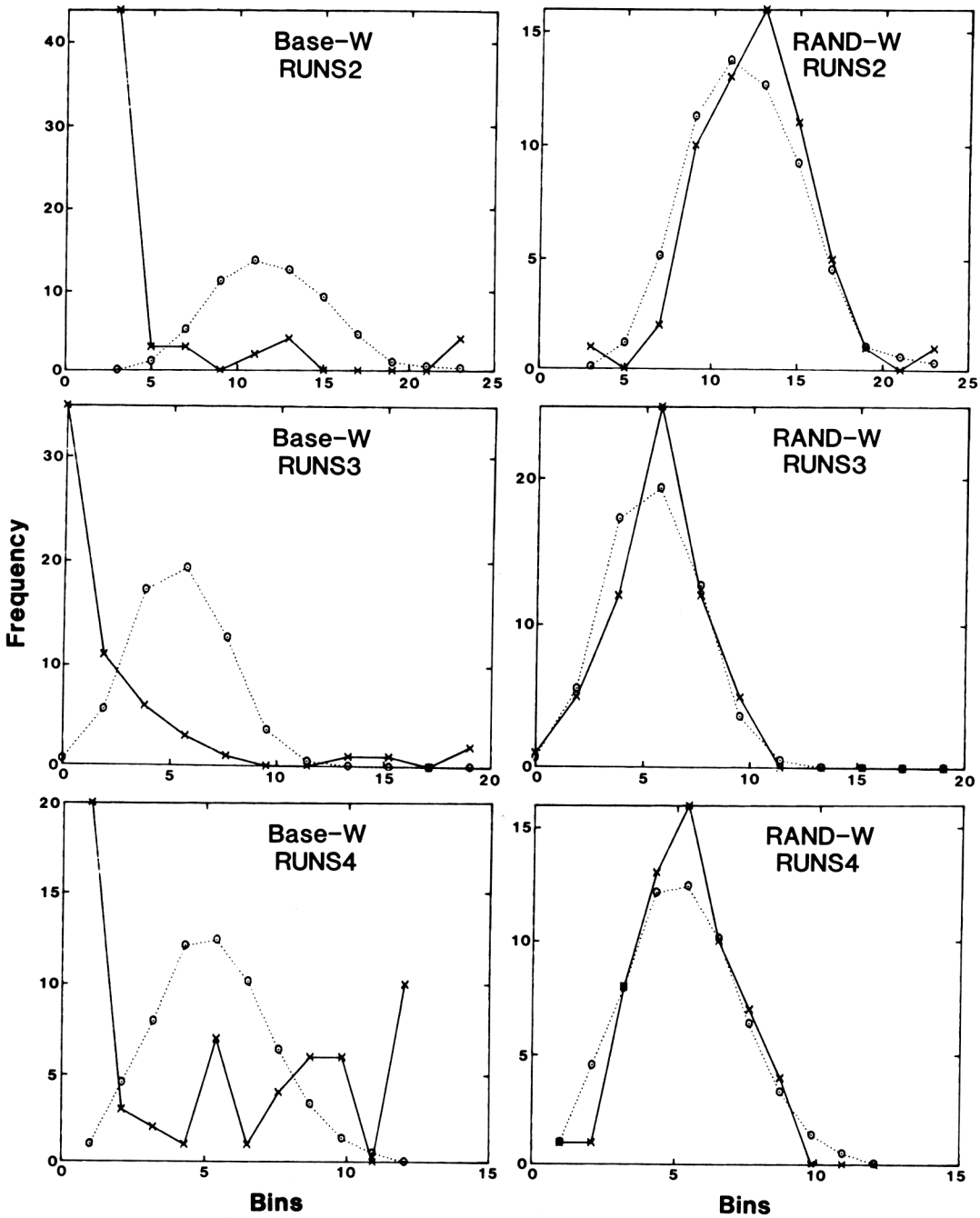
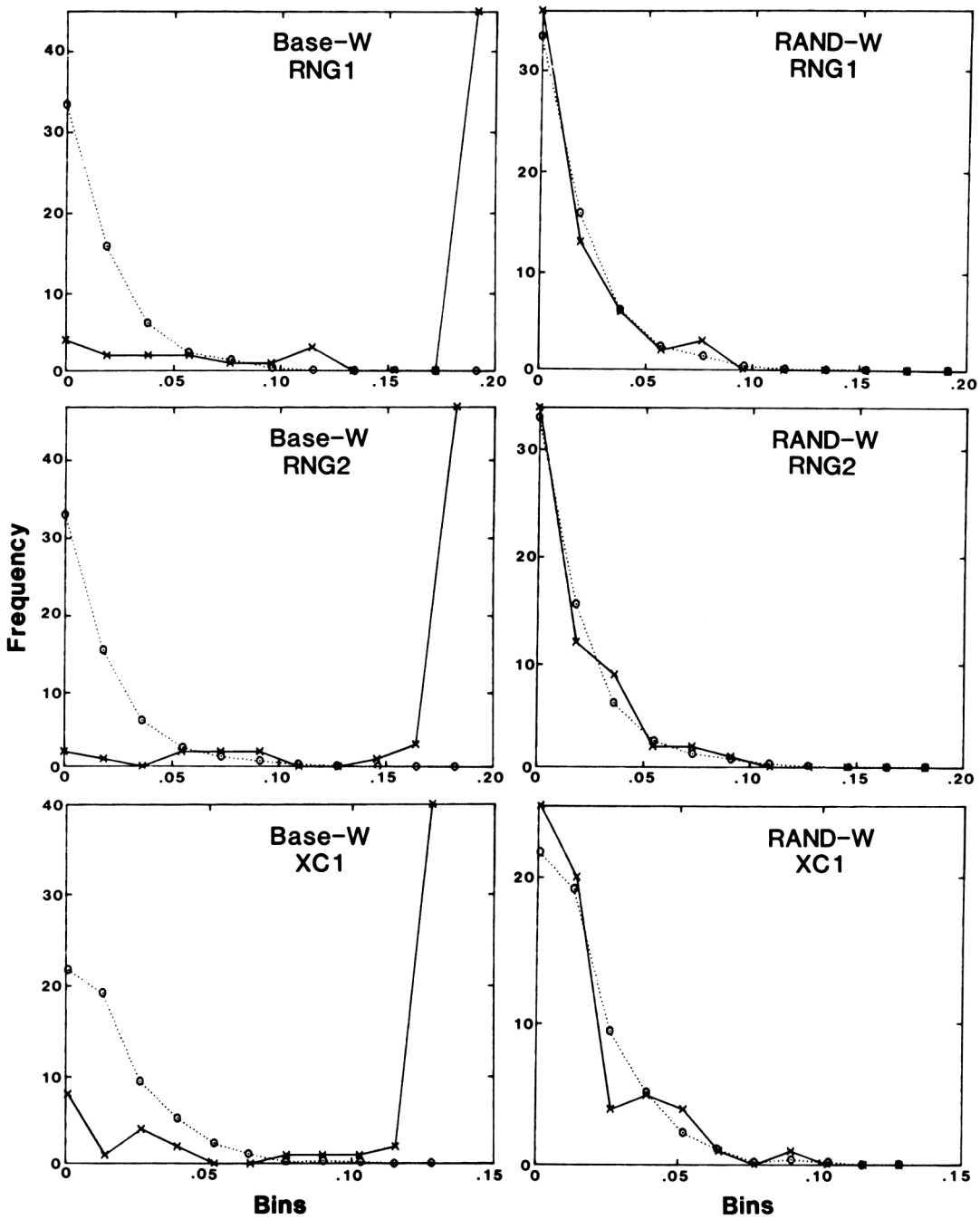


Figure 5. Frequency distributions for Subject W on RUNS2, RUNS3, and RUNS4-10 statistics. (Dotted line shows random number generator.)


Base-W
RNG1
RAND-W
RNG1
Base-W
RNG2
RAND-W
RNG2
Base-W
XC1
RAND-W
XC1
Frequency
Bins
Bins