Behavior Research Methods, Instruments, & Computers
1998, 30 (4), 583-591

Analyzing human random generation behavior: A review of methods used and a computer program for describing performance

JOHN N. TOWSE and DEREK NEIL
Royal Holloway University of London, Egham, England

In this paper, we consider the different methods that have been developed to quantify random generation behavior and incorporate these measurement scales into a Windows95 computer program called RgCalc. RgCalc analyzes the quality of human attempts at random generation and can provide computer-generated, pseudorandom sequences for comparison. The program is designed to be appropriate for the analysis of various types of random generation situations employed in the psychological literature. The different algorithms for the evaluation of a dataset are detailed and an outline of the program is described. Performance measures are available for assessing various aspects of the response distribution, the sequencing of pairs, the ordinal relationships between sets of items, and the tendency to repeat alternatives over different lengths. A factor analysis is used to illustrate the multiple dimensions underlying human randomization processes.

The task of random generation is increasingly becoming part of the psychologist's repertoire for assessing cognitive performance. Random generation data have been used, for example, in the development of theoretical models of cognition (see, e.g., Baddeley, 1986, 1996), for the evaluation of issues in the processing of mental arithmetic among healthy adults (Lemaire, Abdi, & Fayol, 1996; Logie, Gilhooly, & Wynn, 1994), and for exploring the performance of neuropsychological groups, such as patients with Alzheimer's disease (Brugger, Monsch, Salmon, & Butters, 1996) and Parkinson's disease (Robertson, Hazlewood, & Rawson, 1996). Random generation has also been investigated in a developmental context (Rabinowitz, Dunlap, Grant, & Campione, 1989).

The performance of experimental subjects in randomization tasks is commonly assessed by measures of stereotypy (e.g., frequency of adjacent items in an ordinal sequence or preferential selection of particular pairs) and measures of response alternative usage (whether each item in the response vocabulary is equally likely to be selected), although it is also apparent that previous psychological research has incorporated a variety of randomization scores. The absence of a single index of randomization quality is not just a lack of standardized methodology, however. From a logical standpoint, no test can conclusively demonstrate randomness, because randomness itself cannot be directly observed; only departures from randomness (the detection of order) can be quantified. Thus, different measures are required to capture different types of order or pattern that may be found in any sequence. That is, different descriptors are selectively *tuned* to aspects of sequence regularity or predictability. The psychological evidence that shows random generation performance to be a nonunitary trait (see, e.g., Ginsburg & Karpiuk, 1994; see also below) further emphasizes the need for different measures of performance that reflect the efficiency of the separable cognitive components servicing random generation behavior.

However, so far as we can ascertain, there is no current computer software that provides an accurate and flexible analysis of randomization sequences for psychological research. Accordingly, we have developed a VisualBasic program, running under Windows95, that permits different types of randomization responses to be analyzed with several scales in psychological currency and to be printed or graphed as required. Data can be entered directly or imported from a text file and can be archived. In the next section, we review the major tests of randomness reported in the experimental literature, as implemented in the computer program. The measures are presented in conceptual order, according to whether analyses consider responses separately, in relation to another individual response, or for differing sequence lengths.

MEASURES OF RANDOMIZATION PERFORMANCE

Redundancy

In terms of classic information theory analyses (Attneave, 1959; Shannon & Weaver, 1949), a sequence of items can be said to contain maximum first-order infor-

We are grateful to John Valentine for valuable help in the explication of statistical and computational issues and to Roger Bakeman, N. Ginsburg, and an anonymous reviewer for helpful suggestions. A copy of the RgCalc program will be made available from the Internet at the following URL: http://www.pc.rhbnc.ac.uk/cdrg/rgcpage.html. Correspondence concerning this article should be addressed to J. Towse at the Department of Psychology, Royal Holloway University of London, Egham, Surrey TW20 0EX, England (e-mail: j.towse@rhbnc.ac.uk).

mation when each response alternative is selected with equal frequency. As the selection frequency among alternatives deviates from equality, the sequence can be said to have less randomness or more redundancy—hence the name for this measure—since examination of part of the sequence allows for better-than-chance estimations of subsequent choices.

The amount of first-order information that is provided by a sequence is calculated as

$$H_{\text{single}} = \log_2 n - \frac{1}{n}\left(\sum n_i \log_2 n_i\right), \quad (1)$$

where n (here and subsequently) is the number of random responses in the set, and n_i is the number of occurrences of the ith response alternative (computations are omitted where $n_i = 0$). For a response set with a different alternatives, the maximum amount of information that it is possible to generate is

$$H_{\max} = \log_2 a. \quad (2)$$

Accordingly, the redundancy (R) in the sequence is found by determining the extent of deviation from ideal information generation, expressing this value as a percentage score,

$$R = 100 \times \left(1 - \frac{H_{\text{single}}}{H_{\max}}\right). \quad (3)$$

Thus, an R score of 0% indicates no redundancy (perfect equality of response alternative frequencies), and an R score of 100% indicates complete redundancy (the same response choice is used throughout).

Response Frequencies

The production frequency for each response alternative is presented in a tabulated (and, if selected, graphical) form. This provides the opportunity to explore the R score in more detail. It is also possible to request a frequency distribution for digrams (pairs of responses) and trigrams (triplets). In both cases, all combinations that appear at least once are presented, in order of popularity.

Coupon

Ginsburg and Karpiuk (1994) developed a measure of response usage that they termed the coupon score. This measure indicates (across the entire set) the mean number of responses produced before all the response alternatives are given. Accordingly, generation strategies based on *cycling* (working through the set of possible responses) will produce low coupon scores; in randomizing five numbers, the sequence "1, 3, 2, 5, 4" produces the minimum possible coupon score. In the case of a response set where an alternative is omitted entirely, because an exact coupon score cannot be calculated, the program will return a value of $n+$, where n is the sequence length.

Random Number Generation

The redundancy measure described above considers only the distribution of response frequencies; it does not consider the randomness of the *sequence* (the dependency or association between one choice and the next). The random number generation (RNG) score is one popular index of randomization performance that describes the distribution of response pairs, or digrams. The measure was introduced by Evans (1978) and is based on contiguous responses—how often any response alternative follows any other response alternative. Computationally, response pairs are tabulated in an $a \times a$ matrix. Following Evans, in order to produce a full complement of pairs, the final response is paired with the first (thus giving a *wrap-around* function). RNG is then calculated as

$$\text{RNG} = \frac{\sum n_{ij} \log n_{ij}}{\sum n_{ij} \log n_i}, \quad (4)$$

where n_{ij} is the frequency count from each cell in the matrix, and n_i (as described for R above) represents the frequency of occurrence of alternative i. The computation only includes cell values greater than 1, and the final RNG score has a range between 0 (perfect equality of digram distribution) and 1 (complete predictability of pair sequences).

Guttmann's Null-Score Quotient

Brugger et al. (1996) utilized a measure of randomness called the null-score, or NS (Guttmann, 1967, cited in Brugger et al., 1996). NS is the total number of digram permutations that do not appear within the subject response set (again using a wrap-around function to produce the full complement of pairs). This leads to a value between 0 and $a^2 - 1$. To make the NS value more meaningful across different response alternative ranges, the value is expressed here as a percentage of the maximum value attainable—that is,

$$\text{NSQ} = 100 \times \frac{\text{NS}}{a^2 - 1}. \quad (5)$$

From the above description, it should be apparent that the null-score quotient (NSQ) is structurally related to the RNG index, since the former score reflects digrams not used and the latter digrams used repeatedly. Thus, in a reanalysis of data available from 93 subjects who produced written random number sequences (Experiment 1 of Towse & Valentine, 1997), when randomizing among 10 numbers the correlation between RNG and NSQ scores was $r = .97$, $p < .01$. When randomizing among 15 numbers, the correlation between RNG and NSQ was $r = .98$, $p < .01$.

Adjacency

The RNG and NSQ measures consider all possible response pairings, which is often a useful and important

calculation. However, human oral random generation (to take one example) is known to comprise a substantial proportion of a *particular* digram type: adjacent items from the ordinal sequence of alternatives. Thus, letter sequences such as "a, b" (see, e.g., Baddeley, 1966), and number sequences "1, 2" (see, e.g., Wiegersma, 1984a) are common. The *A*, or *adjacency*, measure (sometimes referred to as a stereotyped score) is a more specific or focused measure of digram frequency and is calculated as

$$A = 100 \times \frac{\text{number of adjacent pairs}}{\text{number of response pairs}}. \quad (6)$$

The *A* score is therefore measured as a percentage, and is expressed in the RgCalc program for ascending and descending pairs separately, as well as in the form of a combined value (as for RNG and NSQ scores, a wrap-around function is used to produce the full complement of pairs). *A* scores will range between 0%, in the case where there are no neighboring pairs, and 100%, if the set is entirely composed of such pairs.

Turning Point Index

Azouvi, Jokic, Van Der Linden, Marlier, and Bussel (1996), after Kendall (1976), reported a measure of sequence regularity called the Turning Point Index (TPI). This involves calculating the number of responses that, as numerical values, mark a change between ascending and descending sequences (i.e., points that represent local peaks and troughs in a time-series plot). The number of turning points is compared against a theoretical value

$$\text{TP}_{\text{expected}} = \frac{2}{3}(n-2). \quad (7)$$

Thus, in the sequence "1, 3, 5, 7, 8, 6," there is a single turning point at response "8," as the series begins to descend at this point. In the sequence "5, 3, 4, 6, 2, 8, 9, 7," there are four turning points (on the responses "3," "6," "2," and "9"). Turning points may also straddle response values in the case of repetitions (e.g., the excerpt "2, 4, 4, 3" contains a single turning point between the repeated response "4").

The TPI value is reported as a percentage score, indicating the correspondence between observed and expected values:

$$\text{TPI} = 100 \times \frac{\text{TP}_{\text{observed}}}{\text{TP}_{\text{expected}}}. \quad (8)$$

Thus, values greater than 100% indicate that too many turning points were produced (relative to a theoretical distribution of random responses), whereas values less than 100% indicate fewer turning points than expected. Azouvi et al. (1996) found that patients with closed head injury produced a lower TPI than did controls, indicative of a *runs* strategy where individuals produce an arithmetic *chain* of responses.

TPI values are closely associated with the Wallis–Moore phase frequency test (see Sachs, 1978), which provides a statistic for the persistence of an ordinal trend (i.e., an *absence* of turning points). Given the strong reciprocal relationship, the phase frequency test is not implemented here.

Phase Length

As Kendall (1976) pointed out, it may be informative to consider not only the number of turning points but the distribution of intervals between them. The interval between two turning points is termed a *phase*, so if the third response in a sequence produced a trough and the fourth produced a peak, there would be a phase of 1 between them. A further illustration is given by the sequence "2, 3, 5, 4, 5, 6, 7, 8, 6, 1, 3." There are turning points at the values 5, 4, 8, and 1, and the phase lengths (PLs) between these points are 1, 4, and 2, respectively. The distribution of PLs is calculated for the entire sequence, and the expected frequency of phases with a length or distance *d* is

$$\text{frequency}(d)_{\text{expected}} = \frac{2(n-d-2)(d^2+3d+1)}{(d+3)!}. \quad (9)$$

The PL (expressed as a count value) is the number of observed phases of length *d*, and these scores are presented alongside the expected frequencies from a theoretical random distribution. Accordingly, observed values in excess of expected values indicate that more phases of length *d* were produced than would be predicted in random sequences, whereas values less than expected scores indicate that fewer phases were produced than would be predicted.

It is important to note that the sum of the observed and theoretical values may not be equal (a consideration that essentially prohibits the expression of the PL scores as percentage of observed over expected values). The number of phases that are found in a sequence will depend, in part, on the position of the first and last turning point, since the PLs of the sequence before the first and after the last turning point are unknown. The total number of phases will also depend on the lengths of those phases. A sequence may contain either many short phases or fewer long phases, for any particular number of responses.

Runs

Ginsburg and Karpiuk (1994) describe a measure of randomness called *runs*, which describes the *variability* in the phase lengths (a similar score, ALTS, is described in Neuringer, 1986). From the response set, the number of items in successive ascending sequences is determined (i.e., ascending phase length values). The variance of these sequence lengths is then derived. Ginsburg and

Karpiuk treat repetitions as breaks in the ascending sequence, and this scoring method is adopted here.

First-Order Difference

Tabulated values (and a graphical form of these data) are available for first-order difference (FOD) distributions. This measure (see, e.g., Brugger et al., 1996; Treisman & Faulkner, 1987; Wiegersma, 1984a) reflects the arithmetic difference between each response and its preceding value (as in analyses described above, a wrap-around function is used in calculating the set of values). Thus, the response pair "2, 7" yield an FOD of +5, and the response pair "7, 4" yield an FOD of −3. FODs are calculated for all pairs, and the frequency of each possible FOD permutation is determined. The resulting distribution may point to the arithmetic, or at least ordinal-based, strategy underlying response choices.

It should be apparent that the FOD scores provide a count of adjacent responses (*A* score) as +1 and −1 values. The advantage in these FOD scores is that they illustrate the extent to which adjacent values predominate sets, by describing other permutations also (e.g., counting in twos). Typically, they will also illustrate the avoidance of immediate response repetitions (an FOD value of 0), relative to other digram sequences. In addition, FOD scores are clearly suitable for the graphical presentation of results.

When interpreting FOD distributions against theoretical norms, the expected frequencies derived from random samples will not be linear. There are more permutations of numbers yielding an FOD value of 0 than for any other score. Similarly, +1 and −1 values are more likely than +2 and −2 values in random sets (see Brugger et al., 1996, for the set of permutations on a six-alternative task).

Scores from points along the FOD distribution may be usefully compared across experimental conditions pertaining to randomization (where the number of alternatives is the same). Data might also be analyzed through a one-sample *t* test comparing scores against theoretically expected values (if appropriate, using Hotelling's T^2 to make multiple comparisons while protecting against inflated Type I error rate).

Repetition Distance

As mentioned above, human subjects usually do not repeat response values with a frequency that matches randomly generated sets. However, unless the sequence length is equal to or smaller than the number of possible alternatives, individuals must eventually produce further occurrences of some response choices. Repetition distance data are presented in tabulated (and, if selected, in graphical) form, to show the distribution of distances or lags between item repeats (as used by Zwaan, 1964, and Mittenecker, 1953, both cited in Wagenaar, 1972). Consider the sequence "2, 3, 7, 8, 8, 7, 2, 3, 2." The response "2" is repeated after a lag of six items (i.e., it is the sixth item after the preceding occurrence) and again after a lag of two items. The response "3" repeats after six items also, the response "7" after three items, and the response "8" repeats with a lag of one item. All distances from the entire response sequence are then collated into a repetition distance table.

Towse (1998) presents repetition distances from human performance and also from computer-generated random sets. The latter shows an *approximation* to a geometric distribution of the form

$$\text{frequency}(s)_{\text{expected}} = (1-p)^{s-1} \times p \times (n-1), \quad (10)$$

where s is the number of repetition steps (the lag), and p is the likelihood of an item's selection.

Both theoretically and in human data, the occurrence of repetition distances beyond a lag of approximately 20 items becomes sparse (in the case of 10 response alternatives). Consequently, Towse (1998) presented data in the form of a *bin* for those repetition distances longer than 20 items. However, the point at which a bin of this type becomes useful will depend on the response vocabulary size and the randomization sequence length. Statistical analysis of repetition distance profiles might follow in the same manner as that for FOD scores, indicated above or see next sections.

Repetition Gap

Quantitative measures of repetition performance can be obtained as repetition gap scores (adapting and extending Ginsburg & Karpiuk, 1994). From the table of repetition distances, the mean gap, the median gap, and the modal gap values are determined and displayed.

Phi Index

Wagenaar (1970) and Wiegersma (1984b) provide examples of randomization analysis using a potentially complex measure called a phi coefficient. Since we implement a gain function that potentially produces phi values outside the range −1 to +1, we term this measure a phi index. The phi (ϕ) index is a measure of repetition tendency over different lengths (different orders of analysis) for binary sequences. Nonbinary sets are analyzed by transformation into separate two-alternative sequences for analysis, as described below.

The computation of the ϕ index takes place over several stages. First, we describe the process in broad terms. In essence, the ϕ index shows whether the subject tends to repeat values at a given *d*-gram length, relative to sequence frequencies at shorter lengths. For the analysis of response segments with a distance *d* (i.e., *d*-gram sections), one counts the number of occasions where the first and the last response in all *d*-length sequences are the same (i.e., repeat). Similarly, the number of occasions is counted where the first and last response in all *d*-gram sequences are different (i.e., alternate). These values are then compared with predicted (expected) fre-

quencies for these repeating and alternating strings, based on the known frequency distribution of $(d-1)$-gram sequences. If responses are random over a particular length, therefore, observed and expected frequencies will match, a hypothesis evaluated statistically.

Computationally, the procedure for calculating expected frequencies is

$$\text{frequency}(r_1, r_2, \ldots r_d)_{\text{expected}} = \frac{f(r_i \ldots r_{d-1}) \times f(r_2 \ldots r_d)}{f(r_2 \ldots r_{d-1})}, \quad (11)$$

where r_x is the response for the xth item in the sequence. For two-gram sequences, the denominator is the number of responses that are generated.

The full complement of expected (and observed) frequencies is determined, and then the d-gram lengths are categorized according to whether they represent repeating end points ($r_1 = r_d$) or alternating end points ($r_1 \neq r_d$), and this yields a 2×2 table for the analysis of a given sequence length. χ^2 values are then computed and ϕ derived as

$$\phi = \sqrt{\frac{x^2}{T}} \times 100, \quad (12)$$

where T is the total sequence length (after transformations—see below). The ϕ index, produced separately for sequence lengths or order of analysis of length d, has a potential range between -100 and 100, since a sign is added—a minus value to indicate that more d-grams were alternating than was predicted (negative recency), a plus value to indicate that more d-grams were repetitions than predicted (positive recency). In RgCalc, ϕ is computed for six orders of analysis (for all sequences up to seven items in length).

By way of an example for calculating expected frequencies, consider a binary set of 100 responses with 60 "0" responses and 40 "1" responses. The expected frequencies for the alternating sequence "0,1" and the repeating sequence "0,0" are

$$f(0,1) = \frac{f(0) \times f(1)}{f(0) + f(1)}$$

$$= \frac{60 \times 40}{100}$$

$$= 24$$

and

$$f(0,0) = \frac{f(0) \times f(0)}{f(0) + f(1)}$$

$$= \frac{60 \times 60}{100}$$

$$= 36.$$

Continuing this example for the analysis of one permutation of a three-gram sequence,

$$f(0,0,1) = \frac{f(0,0) \times f(0,1)}{f(0)}$$

$$= \frac{36 \times 24}{60}$$

$$= 14.4.$$

In a case in which a alternatives are randomized and $a > 2$, the sequence must be first translated into a binary sets, where a_n values are recoded as "0" and all other alternatives recoded as "1." This procedure is repeated for all possible alternatives. Values are summed across transformations for observed and expected frequencies to form a summary 2×2 table, although, as a cautionary interpretive point, these successive transformations may not be independent of each other. As the number of alternatives increases, the ϕ score will become smaller, due to increasing prevalence of "1" values in recoded sequences, effectively diluting any repetition bias. This feature, together with variation in the ϕ index according to sequence length, essentially makes comparison of ϕ scores across different experimental parameters meaningful only after some appropriate normalization procedure. Wiegersma (1984b) gives one example of such a normalization treatment. An alternative approach suggested by Wagenaar (1970) is to evaluate all ϕ scores with reference to Monte Carlo simulation scores and obtain relevant percentile rank values.

RNG2 (Analysis of Interleaved Digrams)

Neuringer (1986) reported that, when provided with extensive feedback concerning the statistical adequacy of two-choice, keypress-based, randomization sequences, experimental subjects eventually learned to produce sets that corresponded to computer-generated random numbers, at least as measured by the indices providing feedback. One of the novel randomization descriptors used by Neuringer was a score describing the distribution of interleaved pairs, termed RNG2. This score involves the pairing of every alternate response together to make up a frequency matrix. Thus, in the sequence "2, 3, 7, 8, 8, 7, 2," the digrams are "2, 7," "3, 8," "7, 8," "8, 7," and "8, 2." This produces $n - 2$ digram pairs. The digram pairs, once obtained, are processed just as for the RNG measure.

A DESCRIPTION OF RGCALC

After starting RgCalc, one can bring down the "File" menu to select the "Open/Define Response Alternatives" option. This allows the user to specify the items that form the response vocabulary. This might be a sequence of numbers or letters, but any set can be used, such as the string "1, 2, 5, 6, 9," or nonnumeric cate-

gories, such as letters or days of the week. For convenience, it is also possible to specify a continuous number series as the available response set by entering the required range of values in the "Numerical Series" field and then clicking on the "Generate" button. Response alternatives can be saved or opened as required from the "File" menu option when this window is active.

Once the response set has been specified, the user can close this window in the standard manner (by clicking on the "x" button in the top-right corner) to open a form, or grid, to allow the responses themselves to be entered. This can be done manually, or, if the response set has been prestored as a comma, tab, or return delimited file, the sequence can be imported through the "Open Response File . . ." option from the "File" menu (in fact, any nonprintable character can be used as a delimiter, and data may also be entered to and from the clipboard using the "Cut," "Copy," and "Paste" functions available both from the edit window and by right-clicking on the mouse). When loading a response file, one can select the type of file from the pull-down menu; the default is to list all ".txt" suffix files. It is also possible to specify the desired suffix and then click on the "Open" button to produce a list of all files available in the current directory. Select the desired response file and click on the "Open" button to load this dataset into the program; the status bar will provide an indication of program activity.

When specifying a new set of responses, values are entered, using the response field below the grid. Entering a value and pressing the return key will place that value in the selected response cell and move the field onto the next response cell. Alternatively, one can choose response positions in any order. Point-and-click at a cell, and then enter the appropriate response or use the arrow keys to move around the grid. Note that, initially, the program does not verify that responses are legitimate (i.e., that they are part of the response vocabulary). Entry checking is accomplished when the user attempts to calculate randomization scores by clicking the "Calculate" button, when any inappropriate responses are reported to the user for attention.

Where at least one response has been entered into the response form, the responses can be stored by using the "Save Response File . . ." option from the "File" menu. If the user attempts to close down the response set window after changes have been made to an earlier, saved version of the file, the program will prompt the user to save the latest revision.

Other response entry options are the following:

"Empty Grid" button—all the response cells are emptied to allow a new set of values to be entered. The computer provides a warning if this involves erasing unsaved data.

"Generate" button—takes the value entered in the adjacent field box to specify the number of responses to be generated using an internal, pseudorandom algorithm. This allows an exploration of what an appropriate sequence of random responses might look like according to the various randomness tests.

"Calculate" button—checks that the response set contains only permitted values and then computes the performance descriptors. A progress bar reports on the calculations, and the statistic being computed is displayed (though often for a subthreshold duration). When complete, the results are shown in a new window.

For all the functions above, the user may also press the Alt key together with the relevant, underlined letter on the button name, as a keystroke alternative to mouse operations. For example, Alt plus the C key is equivalent to pressing the "Calculate" button.

Once calculated, random generation scores are displayed on screen or are made available via button selections. By default, the program shows the sample size, along with response frequencies, first-order differences, repetition distances, repetition gap measures, PL, TPI, runs, coupon score, R, RNG, NSQ, and RNG2 scores. Plots of response frequencies, first-order differences, and repetition distances are available, and further information (specification of response pairs and triplets, adjacency scores, and ϕ index values) can be selected.

A "Print" button is provided, which sends the results of specified randomness tests to the default printer. A "Print to file . . ." button allows the user to send (and compile) scores to a computer file for later analysis or for use by other programs. If an already existing file is selected, data are appended to the end of this file. Some results (e.g., preferred response pairs) are not printed, because they contain a variable number of results, making identification of values problematic. Modal repetition gap values are available, but where there is more than a single mode, the first value is given as a negative number, to highlight the presence of other modal values. Scores are saved in the following order: sample size, R, RNG, NSQ, RNG2, TPI, runs, coupon, ascending (adjacent), descending (adjacent), combined (adjacent), response frequencies for each alternative, first-order differences, repetition distance frequency (length 1–20, and a summed value for lengths greater than 20), mean repetition gap, median repetition gap, modal repetition gap, and ϕ index values (orders 2 to 7).

OVERVIEW OF RANDOMIZATION PERFORMANCE MEASURES

In this section, we do not attempt to review the extensive literature on random generation per se (see Brugger, 1997, for an overview). However, we adumbrate the different ways in which human randomization varies according to the experimental conditions under which sequences are generated and, thus, illustrate some of the characteristics of the randomness scales.

Experimental studies have shown that, in oral and written random number generation, the R score is sensitive to the response vocabulary size—that is, the number

of alternatives (see, e.g., Towse & Valentine, 1997). However, the *R* score for random *keypress* tasks does not necessarily vary significantly with set size, which appears to illustrate one difference between generation from an internal set (e.g., numbers) and choice among external referents (e.g., keys). Supporting this view, the provision of response alternatives during random number generation moderates the set size effect (Towse, 1998). Further, some reports suggest that *R* increases for oral random letter generation as response speed increases (Baddeley, 1966), although this does not always appear to be the case for oral number generation (Towse, 1998). The nature of the response set may be relevant, since Baddeley used letters of the alphabet (and, therefore, 26 alternatives). Accordingly, changes in the sequencing of items (e.g., an increasing use of alphabet strings) might impact on the distribution of response usage, as the frequency data are relatively sparse with large response vocabularies.

The production of item associates (particularly neighboring values to the just-articulated item) is strongly related to the temporal interval between responses. The faster the rate of production, the greater the proportion of adjacent items and repetitive pairs that are used (Baddeley, 1966). Thus, measures such as *A* and RNG vary substantially with response speed conditions and, therefore, also with associated scores, such as NSQ, TPI, and PL (since strings of adjacent items are used, the phase length increases). The reliance on adjacent items is also likely to be evident from inspection of the FOD distribution.

Intriguingly, the avoidance of repetitions in random sequences and the repetition distances that are produced seem rather invariant across experimental manipulations (Towse, 1998; Wagenaar, 1970). Wagenaar reported that ϕ varied with the number of alternatives in a keypress task, although it is not clear whether this was due in part to small response vocabularies, since, for example, repetitions are more forced in a two-alternative task. The repetition avoidance phenomenon has been linked to the operation of *competitive queuing* mechanisms in connectionist networks—that is, automatic self-inhibition processes that prevent response perseveration of highly active nodes in a distributed neural network (for a discussion, see Towse, 1998). Repetition avoidance has been reported to be less evident among bilateral hippocampal amnesics (Brugger, Landis, & Regard, 1992).

In sum, it is apparent that, although many of the measures of random generation performance are statistically related to each other (Ginsburg & Karpiuk, 1994; see below), not all scores intercorrelate highly, and not all are sensitive in the same manner. The type of random generation task employed, the speed of response, and the choices available for response, as well as the psychological population under investigation (Brugger et al., 1996) are all important variables. In addition, some characteristic patterns of human random generation, such as the avoidance of repetitions, do not necessarily imply a strategic effort to respond in a particular way; specific phenomena may ultimately be shown to be a by-product of other generation mechanisms. However, unless researchers have at their disposal a variety of performance descriptors, progress on such matters may be slow.

Finally, comparison of randomization statistics must be independent of bias from particular task configurations. Thus, for example, the response length and the number of response alternatives will affect the *baseline* values according to several randomness tests. For example, with more responses, more pairs occur, and, thereby, more repeated pairs are likely (picked up by the RNG measure). Differences in randomization performance across certain manipulations should, therefore, be interpreted in the context of expectations derived from theoretical distributions. One method for achieving this is to use the "pseudorandom set" function of RgCalc, which attempts to produce a pseudorandom sequence from an internal (VisualBasic) algorithm according to the criteria specified by the user. For value-critical Monte Carlo tests, additional sources of random sequences may be sought (see, e.g., Kendall & Babington Smith, 1939).

Measures of Randomness: A Principal Components Factor Analysis

As part of an experiment described in Towse and Valentine (1997), subjects produced written random sequences, using numbers between 1 and 10, inclusive. To explore the measures of randomization available from the RgCalc program, these sequences were reanalyzed and entered into an exploratory principal components analysis. There were 94 subjects in the original corpus of data from a 10-choice condition in Towse and Valentine; 1 individual was dropped from the present analysis as a univariate outlier on a number of variables. Variables that were entered into the principal components model were (in alphabetical order); *A*, coupon, ϕ index(2gram), ϕ(3gram), ϕ(4gram), ϕ(5gram), ϕ(6gram), ϕ(7gram), *R*, repetition gap(mean), repetition gap(median), repetition gap(mode), RNG, RNG2, runs, and TPI. Owing to extreme multicollinearity ($r \geq .95$), NSQ (related to RNG) and phase length(1) (related to TPI) scores were excluded. In those instances in which at least one alternative was not used, a nominal coupon score of 101 was entered, and in cases where there was more than one modal repetition gap, the first lower, value was used.

From examination of the scree plot, four factors were extracted (together, accounting for 66.9% of the variance), and varimax rotation was used, with a cutoff value of .45 for inclusion of a variable in the interpretation of factors (Tabachnick & Fidell, 1996). Table 1 shows the results of this analysis, with variables ordered by (loading) size. Factor labels were constructed on the basis of known statistical properties of the randomness scores, as well as from empirical findings in the psychological literature. Two measures—RNG and repetition gap(me-

Table 1
Variables (Ordered by Size of Loading) Contributing to Factors

Factor 1: Equality of Response Usage	Factor 2: Short Repetitions	Factor 3: Prepotent Associates	Factor 4: Long Repetitions
R	ϕ(2gram)	Runs	ϕ(5gram)
Repetition gap (mean)	ϕ(3gram)	TPI	ϕ(7gram)
Coupon	ϕ(4 gram)	*A*	ϕ(6gram)
RNG2	Repetition gap (mode)	RNG	Repetition gap (median)
RNG			
Repetition gap (median)			

Note—Data are taken from Towse and Valentine (1997).

dian)—substantially contributed to more than one factor; other variables loaded on a single factor only.

The first factor is termed *equality of response usage* and indicates whether individuals use alternatives preferentially (i.e., satisfy the *equipotentiality* criterion in random generation; Towse, 1998). The second factor, *short repetitions*, represents the repetition avoidance tendency over small sequence lengths. The third factor, *prepotent associates*, indicates the tendency to produce stereotyped strings such as adjacent items, whereas the fourth factor, *long repetitions*, represents the repetition tendency over somewhat larger sequence lengths. The complete set of factor loading scores are provided in the Appendix.

We hasten to point out that the principal components analysis, although illuminating, is limited both by the precision and by the comprehensiveness of the dependent variables (i.e., the randomness scores), as well as by the modest sample size for adequate factor analysis. Insofar as the measures lack complete precision in tapping a psychological mechanism, multiple indices that load on each underlying factor might be valuable in psychological analyses, as these tests are likely to be differentially sensitive to performance. With respect to the comprehensiveness constraint, the problem is substantial, in that there are innumerable randomness tests one could implement; additional underlying factors might be established if the appropriate measurement were known. This argument, of course, points to the utility of a psychological analysis of the cognitive operations that underlie human performance, to establish the motivation for developing specific tests to capture some relevant aspect of performance.

In sum, however, despite the limitations in the interpretation of these data, analysis converges strongly with experimental findings in identifying several distinct components to randomly generated sequences and, thereby, extending the factor analysis provided from a smaller sample set and reduced variable range in Ginsburg and Karpiuk (1994).

MEASURES OF RANDOMNESS NOT USED

Although we have examined the literature for different measures of random generation performance, we have not implemented all known or possible randomness tests. A number of measures are variants on those already provided:

Series (Ginsburg & Karpiuk, 1994)—essentially equivalent to the *A* measure.

Repetitions (Ginsburg & Karpiuk, 1994)—measure available from repetition distance table.

Variance of digits (Ginsburg & Karpiuk, 1994)—essentially equivalent to *R* score.

Digram repetitions and cluster ratio (Ginsburg & Karpiuk, 1994)—both a form of RNG score.

Poker (Ginsburg & Karpiuk, 1994)—represents repetitions over an arbitrary sequence length, but essentially available from other measures such as the repetition distance table.

C1 and C2 (Neuringer, 1986)—a measure of the similarity of responses across separate random response sets and, therefore, not computable directly. However, the pairs for this measure can be combined manually and then entered as a single sequence. The RNG and RNG2 score will then equate to C1 and C2 values.

Higher order distributions. Analysis of three-gram and longer sequences is not presented here (except for specific measures such as repetition distances), because human randomization data are generally too sparse to permit appropriate analysis of data.

OTHER FORMS OF THE RANDOMIZATION TASK

Although RgCalc has been designed to be as flexible as possible as an analytic tool, there may be some forms of the randomization task that cannot be dealt with, at least in a straightforward fashion. For example, some researchers have promoted the paradigm of a random interval production task (Stuyven & Van der Goten, 1995). Here, subjects are asked to tap a key and to make the temporal gaps between responses as random as possible. An analogous situation would be the request to make random movements (random lengths) over time. In both cases, subjects produce continuous data rather than choosing between discrete alternative categories. Consequently, analysis of the quality of responses may require more specialized methods (clustering of response values may be revealing, for example, or graphical forms of time-series analyses). However, the response data might also be sorted into an appropriate set of groups

(e.g., temporal intervals or distances). Responses having been transformed into categorical choices, standard analytic tests described above could then be applied.

REFERENCES

ATTNEAVE, F. (1959). *Applications of information theory to psychology*. New York: Holt, Rhinehart & Winston.

AZOUVI, P., JOKIC, C., VAN DER LINDEN, M., MARLIER, N., & BUSSEL, B. (1996). Working memory and supervisory control after severe closed head injury: A study of dual-task performance and random generation. *Journal of Clinical & Experimental Neuropsychology*, **18**, 317-337.

BADDELEY, A. D. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology*, **18**, 119-129.

BADDELEY, A. D. (1986). *Working memory*. Oxford: Oxford University Press, Clarendon Press.

BADDELEY, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology*, **49A**, 5-28.

BRUGGER, P. (1997). Variables that influence the generation of random sequences: An update. *Perceptual & Motor Skills*, **84**, 627-661.

BRUGGER, P., LANDIS, T., & REGARD, M. (1992). The brain as a random generator: The relevance of subjective randomization for neuropsychology. *Journal of Clinical & Experimental Neuropsychology*, **14**, 84.

BRUGGER, P., MONSCH, A. U., SALMON, D. P., & BUTTERS, N. (1996). Random number generation in dementia of the Alzheimer type: A test of frontal executive functions. *Neuropsychologia*, **34**, 97-103.

EVANS, F. J. (1978). Monitoring attention deployment by random number generation: An index to measure subjective randomness. *Bulletin of the Psychonomic Society*, **12**, 35-38.

GINSBURG, N., & KARPIUK, P. (1994). Random generation: Analysis of the responses. *Perceptual & Motor Skills*, **79**, 1059-1067.

KENDALL, M. G. (1976). *Time-series* (2nd ed.). London: Griffin.

KENDALL, M. G., & BABINGTON SMITH, B. (1939). *Tables of random sampling numbers*. Cambridge: Cambridge University Press.

LEMAIRE, P., ABDI, H., & FAYOL, M. (1996). The role of working memory resources in simple cognitive arithmetic. *European Journal of Cognitive Psychology*, **8**, 73-103.

LOGIE, R. H., GILHOOLY, K. J., & WYNN, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition*, **22**, 395-410.

NEURINGER, A. (1986). Can people behave "randomly"?: The role of feedback. *Journal of Experimental Psychology: General*, **115**, 62-75.

RABINOWITZ, F. M., DUNLAP, W. P., GRANT, M. J., & CAMPIONE, J. C. (1989). The rules used by children and adults in attempting to generate random numbers. *Journal of Mathematical Psychology*, **33**, 227-287.

ROBERTSON, C., HAZLEWOOD, R., & RAWSON, M. D. (1996). The effects of Parkinson's disease on the capacity to generate information randomly. *Neuropsychologia*, **34**, 1069-1078.

SACHS, L. (1978). *Applied statistics: A handbook of techniques*. Berlin: Springer-Verlag.

SHANNON, C. E., & WEAVER, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

STUYVEN, E., & VAN DER GOTEN, K. (1995). Stimulus independent thoughts and working memory: The role of the central executive. *Psychologica Belgica*, **35**, 241-251.

TABACHNICK, B. G., & FIDELL, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

TOWSE, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, **89**, 77-101.

TOWSE, J. N., & VALENTINE, J. D. (1997). Random generation of numbers: A search for underlying processes. *European Journal of Cognitive Psychology*, **9**, 381-400.

TREISMAN, M., & FAULKNER, A. (1987). Generation of random sequences by human subjects: Cognitive operations or psychophysical process? *Journal of Experimental Psychology: General*, **116**, 337-355.

WAGENAAR, W. A. (1970). Subjective randomness and the capacity to generate information. *Acta Psychologica*, **33**, 233-242.

WAGENAAR, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, **77**, 65-72.

WIEGERSMA, S. (1984a). Forward and backward continuations in produced number sequences. *Perceptual & Motor Skills*, **58**, 735-741.

WIEGERSMA, S. (1984b). High-speed sequential vocal response production. *Perceptual & Motor Skills*, **59**, 43-50.

APPENDIX
Factor Loading Scores From Principal Components Analysis

Measure	Factor 1	Factor 2	Factor 3	Factor 4
R	**.9451**	.0354	.0565	.0461
Repetition gap (mean)	**−.8272**	−.2364	.0338	−.2283
Coupon	**.7561**	.0796	−.0320	−.1643
RNG2	**.6695**	−.2358	.2701	.1703
ϕ(2gram)	.2107	**.8342**	.0743	.0008
ϕ(3gram)	−.1466	**.8255**	.0124	.0808
ϕ(4gram)	−.0881	**.6693**	−.1044	.2650
Repetition gap (mode)	−.2783	**−.6357**	.0393	−.3268
Runs	−.0905	−.2063	**.8568**	.0243
TPI	−.1466	−.2086	**−.8522**	−.0679
A	.0237	.0322	**.8231**	−.0544
RNG	**.5425**	−.1856	**.6285**	.0634
ϕ((5gram)	−.1121	.1642	.1884	**.7076**
ϕ(7gram)	.0992	.0972	−.0294	**.6654**
ϕ(6gram)	.0677	.0712	−.0321	**.6339**
Repetition gap (median)	**.5243**	−.4028	.0810	**.5834**

Note—Data are taken from Towse and Valentine (1997). Values in bold denote inclusion of variable on a particular factor.

(Manuscript received August 19, 1997;
revision accepted for publication December 1, 1997.)