Physical models as tests of randomness

I. Vattulainen,[1,2] T. Ala-Nissila,[1–3] and K. Kankaala[2,4]

[1] *Research Institute for Theoretical Physics, P.O. Box 9 (Siltavuorenpenger 20 C), FIN-00014 University of Helsinki, Finland*

[2] *Department of Electrical Engineering, Tampere University of Technology, P.O. Box 692, FIN-33101 Tampere, Finland*

[3] *Department of Physics, Box 1843, Brown University, Providence, Rhode Island 02912*

[4] *Centre for Scientific Computing, P.O. Box 405, FIN-02100 Espoo, Finland*

(Received 6 March 1995)

We present and analyze in detail a test bench for random number sequences based on the use of physical models. The first two tests, namely the cluster test and the autocorrelation test, are based on exactly known properties of the two-dimensional Ising model. The other two, the random walk test and the n-block test, are based on random walks on lattices. We have applied these tests to a number of commonly used pseudorandom number generators. The cluster test is shown to be particularly efficient in detecting periodic correlations on bit level, while the autocorrelation, the random walk, and the n-block tests are very well suited for studies of weak correlations in random number sequences. Based on the test results, we demonstrate the reasons behind errors in recent high precision Monte Carlo simulations, and discuss how these could be avoided.

PACS number(s): 02.70.Lq, 75.40.Mg, 05.40.+j, 05.50.+q

I. INTRODUCTION

Since the late 1940s, the Monte Carlo (MC) method [1, 2] has become an established tool in various fields of natural sciences, with applications including MC simulations in physical sciences [3] and stochastic optimization [4] in applied mathematics. The fundamental idea on which the MC method is based is the use of random numbers. For computational purposes random numbers have traditionally been produced by deterministic rules, implemented as pseudorandom number generators (PRNG's) which usually rely on simple arithmetic operations. Although it is obvious that these pseudorandom number sequences can be "random" only in some limited sense, their imitation of random behavior is often good enough for cases in which the quality of random numbers is not an essential requirement. Modern high speed computers and high-precision calculations, however, have caused the requirements for the quality of random number sequences to greatly increase. Under such circumstances it is crucial to confirm the quality of random numbers before using them extensively. In order to carry out this task, tests for randomness are needed.

In the course of time, numerous tests have been suggested (see, e.g., Refs. [5–8], and references therein). Some of these have been constructed to study the properties of PRNG *algorithms* and are therefore theoretical [6, 7]. An alternative approach is to study properties of random number sequences regardless of the source. Such tests are called empirical. Moreover, many of the tests are called *standard* [6] in the sense that they probe purely general statistical properties of random number sequences, not concentrating on the requirements of any application in particular. When correlations are not found, the sequence passes the test.

Passing several tests does not prove the randomness of any sequence, however. This is due to the fact that proving randomness requires that the sequence fulfill an actual *definition* for randomness. An unfortunate fact is, however, that there is *no* unique definition for randomness. Attempting to prove randomness in the case of a (long) sequence of numbers for any of the definitions suggested (see, e.g., Refs. [6, 9–11]) is not feasible. Therefore, the best any test can do is to *build confidence* in the properties of random number sequences. Another practical problem concerns PRNG's. Since they are all based on deterministic algorithms, it is always possible to construct a test for every generator where it will fail. Therefore, passing many tests is never a sufficient condition for the use of any PRNG in all applications. In other words, in addition to standard tests, efficient *application specific tests* of randomness are also needed. This need is emphasized by recent simulations, in which some physical models combined with special algorithms have been found [12–16] which are very sensitive to the quality of random numbers.

During the last two decades, some application specific tests have been proposed and used (for a list see, e.g., Ref. [8]). A systematic test bench has been lacking until now, however. The aim of this work is to present one example of constructing such a test bench following and extending our recent work [8, 17, 18]. We plan to present details of practical implementation of the tests in another publication [19]. The tests have been developed from the point of view of a physicist, in the sense that they are based on direct analogies to physical systems, most notably the two-dimensional Ising model [20] and random walks. Based on these models we have constructed two classes of tests. In the first class based on the Ising model, the *cluster test* [17] compares the cluster size distribution of a random lattice with the Ising model at an infinite temperature. In the *autocorrelation test* [18], we calculate the integrated autocorrelation time of some quantities of the Ising model, when the Wolff updating method [21] is used. The second class comprises tests related to random walks. In the *random walk test* [18], we consider the

distribution of the final position of a random walker on a plane which is divided into four equal blocks. The *n-block test* [18] is based on the idea of renormalizing a sequence of uniformly distributed random numbers, and it is essentially a random walk test in one dimension.

The outline of this paper is as follows. In Sec. II, we first describe the PRNG's used in this work. Following this, in Sec. III we present the tests developed in this work, describing them in detail. A brief account has previously been published in Refs. [17, 18]. The results of the tests are given in Sec. IV. We first demonstrate that a bit level implementation of the cluster test is particularly powerful in finding periodic correlations. Moreover, we show that the autocorrelation, the random walk, and the n-block tests are very efficient in detecting short-range correlations. In particular, we demonstrate that the recent erroneous results obtained in high-precision MC simulations [12–16] are due to these short-range correlations. Our results also support [8] the ideas of Ziff [22], on how the properties of some pseudorandom number generators may be considerably improved. Finally, summary and discussion are given in Sec. V.

II. TESTED PSEUDORANDOM NUMBER GENERATORS

In this section we briefly present the algorithms of the tested pseudorandom number generators. Since most generators are widely used and good reviews of pseudorandom number generation are available [7, 23–25], we will not consider this subject in detail. The only exceptions are the ZIFFp and PENTAp generators, which will be described in some detail due to the lack of published documentation.

The pseudorandom number generators tested in this work include generalized feedback shift-register (GFSR) algorithms GFSR$(p, q, \oplus)$ [26], which are of the form $x_n = x_{n-p} \oplus x_{n-q}$, where $\oplus$ is the bitwise exclusive-or (XOR) operator. They are denoted by Rp; recommended values for p and q $(p > q)$ can be found, e.g., in Refs. [27–31]. Other generators include two linear congruential generators $x_n = (16\,807 \times x_{n-1}) \mathrm{mod}\ (2^{31} - 1)$ [32] known as GGL and $x_{i+1} = (69\,069 \times x_i + 1) \mathrm{mod}\ 2^{32}$ [33] known as RAND, RAN3 [34], which is a lagged Fibonacci generator $x_n = (x_{n-55} - x_{n-24}) \mathrm{mod}\ 2^{31}$, and a combination generator RANMAR [24, 35]. Most generators (excluding the GFSR generators) did not require a special initialization procedure. The GFSR generators were initialized with 32-bit integers produced by GGL. Other initialization methods including the one in Ref. [36] were also checked but the test results were unaffected.

In addition to the generators above, we have tested some new promising generators, which are based on the GFSR method with four lags. Their algorithm is GFSR$(p, q_1, q_2, q_3, \oplus)$ or

$$x_i = x_{i-p} \oplus x_{i-q_1} \oplus x_{i-q_2} \oplus x_{i-q_3}, \tag{1}$$

in which $p > \max(q_1, q_2, q_3)$. For the choice of lags, there are two possible approaches. Kurita and Matsumoto [29] have suggested lags based on the theory of primitive pentanomials. In this work, such generators will be called PENTAp generators. Ziff [22] has used a rather different approach, developing generators based on k decimation (i.e., only every kth number of the sequence is used) of GFSR$(p, q, \oplus)$ generators with some value of k which is not a power of 2 (such as $k = 3, 5, 7$). The theory underlying the choice of lags p, q_1, q_2, and q_3 in Ziff's method is given in Ref. [22]. In this work, generators of this kind will be called ZIFFp generators. One particular generator has been given in Ref. [37]. The initialization of PENTAp and ZIFFp generators was performed bit by bit by using GGL: all $b \times p$ bits (b being the word length) in the initial seed vector were initialized by using the most significant bits of integers produced by GGL.

III. PRESENTATION OF TESTS

In the following, we give a detailed account of the new tests. The first two, the cluster test and the autocorrelation test, are closely related to the two-dimensional Ising model [20]. These tests are based on studies of the cluster size distribution in a random lattice, and on calculations of the integrated autocorrelation times for certain physical quantities, respectively. Although we apply these tests to the Ising model, they can be generalized to other models and applications as well. For example, although our version of the cluster test is implemented for studies of random bits, its use for testing random words is a trivial extension. Moreover, the idea of using autocorrelation functions in testing of random numbers is universal.

The next two test methods are related to random walks. In the random walk test, we study random walks on a plane as a function of the walk length. The n-block test is based on the idea of renormalizing a sequence of uniformly distributed random numbers, and is basically a random walk test in one dimension. Despite its simplicity, the latter test is especially effective in finding short-range correlations. In connection with these two tests, we also use the well-known chi-square test [6, 38].

A. Tests based on the Ising model

1. *Cluster test*

There is a natural analogy between binary numbers and the Ising model, which can be made use of in constructing a *cluster test* [17] in the following way. We take the ith bit from every successive number and put them on a two-dimensional square lattice of size L^2. By identifying ones and zeros with the "up" and "down" spins $\mathcal{S} = \pm 1$ of the Ising model, the resulting random configuration should be one of the 2^{L^2} equally weighted configurations corresponding to infinite effective temperature. The simplest quantity that characterizes order in the Ising model is the average magnetization

$$m \equiv \frac{1}{L^2} \sum_{i=1}^{L^2} \mathcal{S}_i, \tag{2}$$

in which $\mathcal{S}_i$ is the spin at site i. In the disordered phase,

the configuration average $\langle m \rangle = 0$. This simple quantity tests the equidistribution of bits.

However, a better measure of *spatial* correlations between the spins can be obtained if we study the distribution of connected spins, or clusters of size s on the lattice. The cluster size distribution $\langle \mathcal{C}_s \rangle$ is given by [39]

$$\langle \mathcal{C}_s \rangle = s p^s D_s(p), \tag{3}$$

in which $D_s(p)$'s are polynomials in $p = 1/2$. The normalization condition is $\sum_{s=1}^{\infty} \langle \mathcal{C}_s \rangle = 1$. Enumeration of the polynomials $D_s(p)$ is a difficult combinatorial problem, and has been done up to $s = 17$ [39]. They are listed in a suitable form, e.g., in Ref. [8].

We note that a similar approach could be utilized in one dimension also. There, the exact solution for the cluster size distribution is known [40], which makes it possible to develop a more complete test (for any s) based on the same physical quantity as in the cluster test. However, the two-dimensional case is more interesting from the point of view of MC simulations.

The test procedure we have used is as follows. We first form an L^2 lattice as above and enumerate all the clusters in it [41] by using periodic boundary conditions in both directions ([42], pp. 26–28). For such a configuration we calculate the (unnormalized) average size of clusters within $s = 1, 2, \ldots, 17$, denoted as $S_{17}^{(k)}$. This procedure is repeated N times corresponding to configurational averaging, yielding $S_{17} = \sum_{k=1}^{N} S_{17}^{(k)}/N$. The theoretical counterpart of this quantity is given by $s_{17} = \sum_{s=1}^{17} s\langle \mathcal{C}_s \rangle$. We also compute the empirical standard deviation σ_{17} of the quantities $S_{17}^{(k)}$. For each ith bit the test statistic chosen is

$$g_i = \frac{S_{17} - s_{17}}{\sigma_{17}}. \tag{4}$$

Using this statistic, the tests were performed comparatively between several pseudorandom number generators, with results from GGL assumed to be independent variables. Comparison of other generators with GGL is justified since GGL has been shown to have excellent properties on bit level [43, 44]. Furthermore, in Fig. 1 we show the distribution for $S_{17}^{(k)} - s_{17}$ as computed for GGL, which is accurately given by a Gaussian.

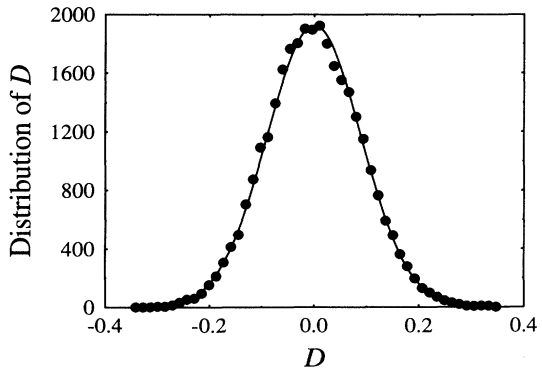

FIG. 1. A schematic (unnormalized) distribution (circles) for GGL of $D = S_{17}^{(k)} - s_{17}$ with 31 000 independent samples. The distribution approaches Gaussian behavior (solid line).

For the final test statistic, the mean value of g_i averaged over all the 31 bits of GGL, denoted as g_{GGL}, and the corresponding standard deviation σ_{GGL} were computed and the results for all other generators were compared to these values using

$$g_i' = \frac{|g_i - g_{\mathrm{GGL}}|}{\sigma_{\mathrm{GGL}}}. \tag{5}$$

The bit i in question failed the test if g_i' was consecutively greater than 3 in two separate tests.

We also considered other similar choices for the test parameters such as using the maximum value of g_i over all the 31 bits of GGL instead of g_{GGL}, and then performing the analysis as above. The results of this approach were consistent with Eq. (5) (results for bits 7 and 12 of RAND being the only exceptions).

2. *Autocorrelation test*

In the autocorrelation test [18] we consider the autocorrelation function C_A for some physical quantity A. Then, we calculate the integrated autocorrelation time τ_A of C_A. Our approach follows the procedure given in Ref. [45].

The autocorrelation function is defined as

$$C_A(t) = \frac{\langle A(t_0)A(t_0+t)\rangle - \langle A(t_0)\rangle^2}{\langle A(t_0)^2\rangle - \langle A(t_0)\rangle^2}, \tag{6}$$

in which t denotes time. In order to calculate an estimator $\tau_A(W)$ for the integrated autocorrelation time τ_A, a truncation window W is used:

$$\tau_A(W) = \frac{1}{2} + \sum_{t=1}^{W-1} C_A(t) + R(W), \tag{7}$$

with the remainder

$$R(W) = \frac{C_A(W)}{1 - \gamma(W)} \tag{8}$$

and

$$\gamma(W) = \frac{C_A(W)}{C_A(W-1)}. \tag{9}$$

The convergence of $\tau_A(W)$ must be checked as a function of the window size W. Since noisy contributions from large separations appear after some value W_n, the estimate τ_A is found by averaging $\tau_A(W)$ between W_c and W_n, in which $W_c < W_n$ denotes the value for which Eq. (7) first converges. An illustration of this procedure is given in Fig. 2. The error estimate for $\tau_A(W)$ is given in Ref. [45].

In this work, we consider the two-dimensional Ising model at its critical coupling point K_c. Since the correlation length associated with the model diverges at K_c, we expect the system to be particularly sensitive to additional spatial correlations due to random numbers during MC simulations. The simulations were carried out

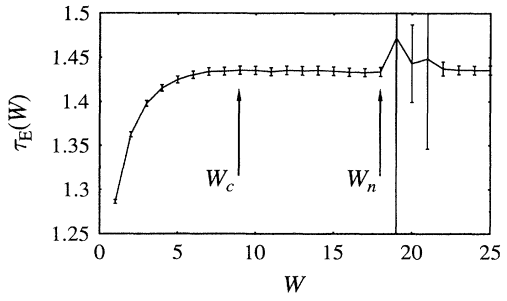

FIG. 2. The integrated autocorrelation time τ_E of energy E, when RAND has been employed. The error starts to increase after $W_n = 18$. The error bars in the cases of $W = 19$ and $W = 21$ extend beyond the boundaries of the graph.

on a square lattice with the Wolff algorithm [21] using $K_c = \frac{1}{2}\ln(1+\sqrt{2})$. The linear size of the system was $L = 16$. Our implementation of the single cluster search algorithm followed Ref. [41], and the measurements for the calculated quantities included the energy E, the magnetic susceptibility $\hat{\chi}$ [45], and the (normalized) size of the flipped clusters c, separated by a single cluster update. Then, by following the procedure given above we calculated the corresponding *integrated autocorrelation times* $\tilde{\tau}_E$, $\tilde{\tau}_{\hat{\chi}}$, and $\tilde{\tau}_c$ from the autocorrelation functions $C_E, C_{\hat{\chi}}$, and C_c. Finally, the estimates for the integrated autocorrelation times were scaled to the time unit of one MC step; i.e., every spin on the lattice is updated once on the average. Therefore the final results are $\tau_A = \tilde{\tau}_A\langle c\rangle$ [45], in which A is one of the quantities E, $\hat{\chi}$, and c.

In the case of the Ising model, the exact value for the energy $E = 1.453\,12$ [46] is known, which allows a comparison between results from different pseudorandom number generators. For other quantities, the test provides us with information on the relative performance of the random number generators. Here we assumed the results from GGL and RANMAR to be correct. This assumption is justified because their results for the energy E were correct within our error limits.

B. Tests based on random walks

1. *Random walk test*

In the random walk test [18] we consider random walks on a two-dimensional lattice, which is divided into four equal blocks, each of which has an equal probability to contain the random walker after a walk of length n. The test is performed N times, and the number of occurrences in each of the four blocks is compared with the expected value of $N/4$, using the chi-square test with three degrees of freedom. A generator fails if the χ^2 value exceeds 7.815 in at least two out of three independent runs [47]. This should occur with a probability of only about 3/400.

In addition to the measure given above, other quantities may also be studied. For example, one may follow the probability distribution function (PDF) on the position of a random walker as a function of time. Comparing such functions between different random number generators may give further information on their relative properties. Calculation of the spatial distribution for the second moment is then also possible. These studies are beyond the scope of the present work, however.

For the purpose of completeness, let us mention that other random walk tests have been proposed by Binder and Heermann ([42], pp. 76–80) and Ziff [22]. The former is based on the idea of studying the average end-to-end distance which should be a linear function of the walk length n. The test proposed by Ziff is based on random walks in a two-dimensional square lattice, where the random walker starts from one corner and heads towards the opposite one. At every step it may turn either left or right, unless it enters a previously visited site, in which case it is forced to turn so as not to retrace its path. Therefore eventually it hits one of the two opposite boundaries, which should occur with an equal probability.

2. *n-block test*

The n-block test [18] is a simplified version of the random walk test, being basically a random walk test in one dimension. In this test we take a sequence $\{x_1, x_2, \ldots, x_n\}$ of uniformly distributed random numbers $0 \le x_i < 1$, whose average $\bar{x}$ is calculated. If $\bar{x} \ge 1/2$, we choose $y_i = 1$; otherwise $y_i = 0$. This is repeated N times. We then perform the chi-square test on variables y_i with one degree of freedom. Each test is repeated three times, and the generator fails the test with fixed n if at least two out of three χ^2 values exceed 3.841 [47], which should occur with a probability of about 3/400.

We note that the main difference between the n-block test and a one-dimensional random walk is in the use of random numbers. In the n-block test random numbers are added together in blocks, and therefore properties of all bits are studied. In the one-dimensional random walk the situation is a little bit different, since the most significant bit is the only one that matters: at each step, the most significant bit determines the direction of the jump.

Finally, let us mention that in Ref. [15] Grassberger has proposed a somewhat unspecified "block" test to study the range of correlations for the lagged Fibonacci generator (LF) $(17, 5, +)$.

IV. RESULTS

A. Tests based on the Ising model

1. *Cluster test*

We have implemented the cluster test to study bit level correlations. Each bit of the random number generators was subjected to the test, and the results were compared with previous results [44] of the d-tuple and rank tests [35].

We first tested the average magnetization, i.e., the equidistribution of bits. The bits failed the test if the

deviation from the expected number of ones (i.e., $L^2/2$) consecutively exceeded three times the standard deviation of the binomial distribution with M samples. The test was repeated twice with $M = 4 \times 10^8$ and its results are also shown in Table I. No correlations were found for GGL, R250, or R1279. Surprisingly, however, this rather simple test revealed that the first 11 bits of RAN3 fail (with standard deviations larger than 6.7) although only the first four or five bits fail in the other tests. On the other hand, for RAND only bits $22 - 31$ failed, which produced an exact 50-50 distribution of zeros and ones.

The actual cluster test for $1 \leq s \leq 17$ was repeated twice with parameters $L = 200$ and $N = 10\,000$. To check for finite-size effects additional tests with $L = 500$ were carried out. They gave results fully consistent with $L = 200$. Our results are summarized in Table I, where results of the previous d-tuple and rank tests from Ref. [44] have also been included. More detailed results of the cluster test (values for test statistics g_i' for all i and for all generators) are given in Ref. [8].

Although more powerful than the other methods, the cluster test reveals no discernible correlations for either GGL, R250, or R1279. For RANMAR and RAN3, the cluster test gives results consistent with Ref. [44], but for RAND additional correlations are revealed in bits $7 - 12$, in contrast to passing the d-tuple test [44]. According to the results of RAND, the cluster test is very effective in locating periodic correlations, since the period of bit number 8 of RAND is as large as 2^{24} [48].

In conclusion, the cluster test in the form presented here is very well suited to detection of correlations on bit level, being especially effective for periodic correlations as shown in detail in Refs. [8, 17].

2. *Autocorrelation test*

The autocorrelation test was carried out with two sets of parameters. First, 10 000 Monte Carlo steps (MCS's) were performed to equilibrate the system starting from a random initial state, and then $N = 10^7$ samples were taken to test most of the generators once. One MCS denotes updating of each lattice site once on the average. In the second set, 100 000 MCS's were followed by $N = 10^8$ samples to test some of the generators more extensively. The linear size of the system was $L = 16$.

A summary of the results in Table II shows that, based on this test, the generators can be classified into two categories. First, let us consider results with $N = 10^7$ samples. For the energy $\langle E \rangle$, deviations from the exact result of $\langle E \rangle = 1.453\,12$ [46] for R31, R250, R521, and RAN3 are much larger than 3σ in which σ is the standard deviation [45]. In particular, the average size of flipped clusters $\langle c \rangle$ is very sensitive to correlations in random number sequences, since in the erroneous cases it is clearly biased. This is illustrated in Fig. 3, in which the PDF's of the flipped cluster size c in the case of few random number generators are given. Most striking, however, is the behavior of the integrated autocorrelation times τ. For generators, which show no significant deviations in $\langle E \rangle$, $\langle \hat{\chi} \rangle$, or $\langle c \rangle$, results for the τ's agree well with each other. However, for R31 and R250, the integrated autocorrelation times show errors of about 8% compared with results of GGL and RANMAR. Our results thus show that these quantities are particularly sensitive measures of correlations in pseudorandom number sequences.

Another important point is the behavior of R31 compared with ZIFF31 and PENTA31. Though R31 clearly fails these autocorrelation tests, its 5-decimated sequence ZIFF31 and a generator PENTA31 based on a primitive pentanomial $x^{31} + x^{23} + x^{11} + x^9 + 1$ give correct results within error limits. This is further emphasized by studies with $N = 10^8$ samples, where R89 fails whereas ZIFF89 and PENTA89 give results as good as RANMAR. Therefore these results clearly indicate that k decimation of GFSR generators with two lags and primitive pentanomials generate sequences with less discernible correlations than GFSR generators based on two lags only.

To compare our results with those of Refs. [13, 16] we also used the autocorrelation time test to further study the decimation of the output of R250, i.e., we took every kth number of the pseudorandom number sequence. For $k = \{3, 5, 6, 7, 9, 10, 11, 12, 24, 48\}$, the correlations vanish in agreement with Ref. [13] ($k = 5$) and Ref. [16] ($k = 3, 5$). On the other hand, for $k = 2^m$ with $m = \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$, the sequences fail. These findings agree with the theoretical result of Golomb [49], who showed that the decimation of a maximum-length GFSR sequence by powers of 2 results in equivalent sequences.

Our results for the autocorrelation test are in agreement with observations made by various other authors

TABLE I. Results of the cluster test ($k = 1$), where bit number one denotes the most significant bit. d-tuple and rank test results are from Ref. [44]. The last column denotes bits which fail in testing the equidistribution of bits.

Random number generator	Failing bits			
	Cluster test	d-tuple test	Rank test	Equidistribution of bits
GGL	None	None	None	None
R250	None	None	None	None
R1279	None	None	None	None
RANMAR	$25 - 31$	$25 - 31$	$25 - 31$	$25 - 31$
RAN3	$1 - 4$, $25 - 30$	$1 - 5$, $25 - 30$	$1 - 5$, $26 - 30$	$1 - 11$, $24 - 30$
RAND	$7 - 31$	$13 - 31$	$18 - 31$	$22 - 31$

TABLE II. Results of simulations for the Ising model at criticality with the Wolff algorithm. The number of samples is denoted by N and the values of lags q_i are given where needed. The value of the decimation parameter k is one unless stated otherwise. The average size of the flipped clusters is normalized by the system size. The errors shown correspond to σ [45]; the most erroneous results are in boldface. See text for details.

N	RNG	k	Lags	Physical quantities			Integrated autocorrelation times		
			q_i	$\langle E \rangle$	$\langle \hat{\chi} \rangle$	$\langle c \rangle$	τ_E	$\tau_{\hat{\chi}}$	τ_c
10^7	R31		3	**1.46774(7)**	**0.564(2)**	**0.5664(3)**	**1.233(4)**	**1.058(3)**	**0.507(2)**
	PENTA31		23,11,9	1.45300(10)	0.545(2)	0.5454(2)	1.440(6)	1.224(5)	0.627(7)
	ZIFF31		13,8,3	1.45313(10)	0.545(2)	0.5456(2)	1.440(6)	1.223(5)	0.624(4)
	R250		103	**1.45509(7)**	0.548(2)	**0.5474(2)**	**1.333(4)**	**1.143(4)**	**0.589(4)**
		3	103	1.45302(7)	0.545(2)	0.5452(2)	1.446(6)	1.226(5)	0.628(5)
	R521		168	**1.45379(7)**	0.546(2)	**0.5461(2)**	**1.384(5)**	**1.182(5)**	**0.604(4)**
	R1279		418	1.45312(7)	0.545(2)	0.5454(2)	1.426(5)	1.215(4)	0.622(3)
	R2281		1029	1.45311(7)	0.545(2)	0.5456(2)	1.439(5)	1.226(5)	0.627(5)
	R4423		2098	1.45303(7)	0.545(2)	0.5454(2)	1.441(5)	1.226(5)	0.624(4)
	R9689		4187	1.45313(7)	0.546(2)	0.5455(2)	1.444(5)	1.229(5)	0.625(4)
	R19937		9842	1.45204(7)	0.545(2)	0.5452(2)	1.434(5)	1.220(5)	0.624(4)
	R44497		21034	1.45292(7)	0.545(2)	0.5452(2)	1.434(5)	1.219(5)	0.622(2)
10^7	RAN3			**1.45254(7)**	0.545(2)	0.5446(2)	1.447(5)	1.231(5)	0.630(3)
	RAND			1.45304(10)	0.545(2)	0.5454(2)	1.434(5)	1.221(5)	0.620(2)
	GGL			1.45309(7)	0.545(2)	0.5454(2)	1.436(5)	1.221(5)	0.622(4)
	RANMAR			1.45303(7)	0.545(2)	0.5452(2)	1.443(5)	1.227(5)	0.624(4)
10^8	R89		38	**1.45720(3)**	**0.5503(6)**	0.55031(4)	**1.3041(14)**	**1.1134(12)**	**0.5662(9)**
	PENTA89		69,40,20	1.45300(3)	0.5454(6)	0.54532(4)	1.4454(16)	1.2278(15)	0.6260(12)
	ZIFF89		61,38,33	1.45305(3)	0.5454(6)	0.54539(4)	1.4398(19)	1.2243(17)	0.6241(15)
	RANMAR			1.45304(3)	0.5454(6)	0.54539(4)	1.4372(16)	1.2211(14)	0.6231(14)

[12, 13, 16], who have also studied the two-dimensional Ising model. Moreover, the errors in the average cluster sizes formed with the Wolff algorithm show that the origin of errors observed in these references lies in local correlations present in the cluster formation process. The main advantage of our approach is the use of integrated autocorrelation times as measures for correlations since the errors are as large as of the order of several percent, whereas for other quantities such as the energy the error is much smaller. Due to the fact that this test is not restricted to the Ising model only, its use in other problems might also prove very fruitful.

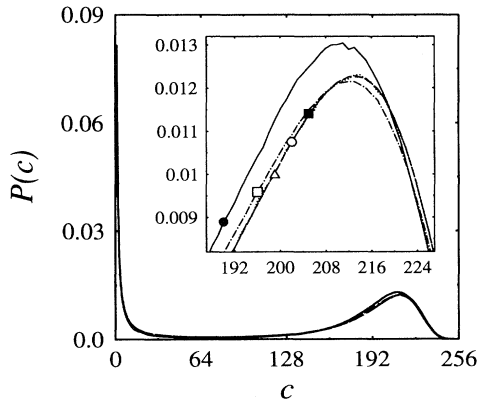

FIG. 3. The (normalized) probability distribution function $P(c)$ of the flipped cluster size c for some random number generators. In the inner figure, a part of the distributions has been magnified. Results of R31 (full circle) and GGL (full square) correspond to $N = 10^7$ samples. In the cases of R89 (open square), ZIFF89 (open triangle), and RANMAR (open circle) a number of 10^8 samples have been taken. Error bars are smaller than the size of the symbols. Results for R31 and R89 clearly deviate from the others.

B. Tests based on random walks

1. Random walk test

Errors in the average cluster sizes for some of the GFSR generators in the autocorrelation test suggest that the correlations must be within the $\mathcal{O}(L^2)$ successive pseudorandom numbers used in the cluster formation. This result is in qualitative agreement with the idea that for GFSR generators the dominant correlations are of triple-point type [15, 22, 50] and thus separated by the longer lag p in the algorithm. To quantify the range of correlations empirically, we have employed the random walk test in a systematic fashion.

First, we studied a group of generators with the walk length $n = 1000$. These results are presented in Table III, and they are in agreement with the autocorrelation test. No correlations for GGL, RAND, or RANMAR were observed. R250 and R521 pass the test with $k = 3$, but fail with $k = \{1, 2, 2^6\}$, whereas R1279 passes with all k's tested. The failure of RAN3 with $k = 1$ is consistent with results of previous tests [44] and the autocorrelation test (however, RAN3 passed the test when every second or third number was used). It is notable that all the failures in this test were very clear, since even the smallest χ^2 values exceeded 40.

TABLE III. Results of the random walk test with $N = 10^6$ samples. The parameter k equals one unless stated otherwise. The fourth column indicates the χ^2 values in three independent tests, or the range of the χ^2 values when results with more than one value of k are included on the same line. The classification of the generators is based on the failing criterion given in Section III B 1: a generator fails the test if the χ^2 value exceeds 7.815 in at least two out of three independent runs. See text for details.

RNG	k	q_i	χ^2 values	Result
R31		3	4094, 4105, 4300	FAIL
R250	1,2,64	103	396.4 – 539.8	FAIL
R521	1,2,64	168	49.01 – 79.16	FAIL
RAN3			40.01, 42.99, 44.53	FAIL
R250	3	103	0.301, 0.873, 1.024	PASS
R521	3	168	1.249, 1.352, 1.735	PASS
R1279	1,2,3,64	418	0.709 – 9.372	PASS
R4423		2098	0.621, 1.226, 8.217	PASS
PENTA31		23,11,9	0.685, 1.587, 2.363	PASS
ZIFF31		13,8,3	2.352, 2.367, 2.632	PASS
RAND			0.304, 0.640, 4.063	PASS
RAN3	2,3		0.033 – 6.877	PASS
GGL			0.090, 0.459, 1.981	PASS
RANMAR			0.293, 1.944, 3.187	PASS

The main difference between the failing generators, R250 and R521 (with $k = 1$), and the successful ones, R1279 and R4423 lies in the lag parameter p, which is less than n for the former and larger than n for the latter. We studied this systematically for various values of p with the random walk test by locating the approximate value n_c, above which the generators fail. The test was performed for R31, R250, R521, and R1279 with $N = 10^6$ samples. The results for n_c were 32 ± 1, 280 ± 5, 590 ± 5, and 1515 ± 5, respectively, in which the error estimate is the largest distance between samples close to n_c. For the purpose of illustration, in Fig. 4 we show an example of the χ^2 values for R31 and R250 as a function of the walk length n.

We also studied GFSR generators with four lags. As Table III indicates, PENTA31 and ZIFF31 pass the random walk test with $N = 10^6$ samples. In these cases, studies to locate n_c were inconclusive, since a small period of these generators did not allow testing them with more than 10^7 samples. Therefore, similar studies for PENTA89 and ZIFF89 with $N = 10^8$ samples were carried out, these results being summarized in Table IV. Although large fluctuations are still present, we may notice that both PENTA89 and ZIFF89 exhibit correlations for $n_c \approx 95 - 200$.

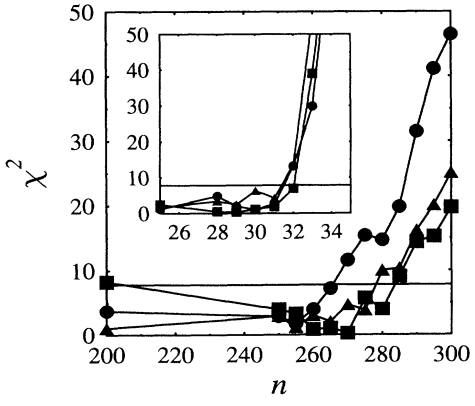

FIG. 4. The χ^2 values for R31 (inner figure) and R250 ($k = 1$) in the random walk test as a function of walk length n, when $N = 10^6$ samples have been taken. Three independent runs in both cases are denoted by different symbols. The horizontal lines denote $\chi^2 = 7.815$.

2. *n*-block test

In the n-block test we used an approach similar to the random walk test. First, we studied various generators with parameters $n = 10^4$ and $N = 10^6$. In the cases of GGL, RAND, RANMAR, and RAN3, we observed no correlations. Studies with RANMAR were repeated with parameters $n = 5000$ and $N = 10^8$, but still no correlations were observed. Then, for GFSR generators R31, R250, and R521 we performed an iterative study by varying n. When $N = 10^6$ samples were taken, the resulting correlation lengths n_c were 32 ± 1, 267 ± 5, and 555 ± 5, respectively. With better statistics, $N = 10^8$, we observed no change for R31, whereas the estimate for R521 reduced to 525 ± 1, and that of R250 to 251 ± 1. The latter value was confirmed with $N = 10^9$ also. Typical values of χ^2 for R250 are shown in Fig. 5, where a sharp onset of correlations at n_c is visible.

Following this, we concentrated on studying the generator ZIFF9689 [GFSR(9689,471,314,157,⊕)] [22, 37], which unlike several other generators has performed well in recent simulations of self-avoiding random walks [14, 15]. This generator was extensively tested up to $n = 25\,000$ and $N = 10^7$, but no correlations were found. In order to increase the number of samples N, we tested ZIFF1279 [GFSR(1279,598,299,216,⊕)], which is a 5 decimation of GFSR(1279,216,⊕). With parameters up to

TABLE IV. Some results of the random walk test with $N = 10^8$ samples for PENTA89 and ZIFF89. For both generators three independent runs have been performed. Failing results are shown with bold type. See text for details.

n	χ^2 for ZIFF89			χ^2 for PENTA89		
85	7.785	7.544	**8.131**	0.427	1.132	1.711
90	2.050	3.716	2.165	1.338	3.874	1.509
95	**10.93**	6.642	3.895	1.335	3.563	3.422
100	**9.130**	5.910	**8.770**	3.867	5.611	4.350
200	**8.632**	**15.76**	**13.61**	**25.02**	**18.48**	**17.56**
500	1.007	6.173	**9.822**	**34.90**	**39.74**	**39.51**

$n = 1500$ and $N = 10^9$, no correlations were observed. These results suggest that deviations from random behavior are much less significant for ZIFF generators than for GFSR generators with two lags. For quantitative purposes, we have studied this subject in more detail by comparing the results of R89, PENTA89, and ZIFF89. These results are shown in Fig. 6. Figure 6(a) clearly shows how dramatically inferior GFSR generators based on primitive trinomials are when compared with generators which are based on either decimation of such sequences or use of primitive pentanomials. Furthermore, when PENTA89 and ZIFF89 are compared with each other with higher statistics $N = 10^9$ [Fig. 6(b)], we may notice that at least in this particular case the decimated sequence ZIFF89 performs somewhat better than PENTA89, although correlations in both sequences are now clearly present.

The results of the random walk and n-block tests show that they are very powerful in detecting rather weak correlations in random number sequences. As far as the generators are concerned, for GFSR generators with two lags the origin of the errors in the simulations presented here and in Refs. [12–16] must be the appearance of rather short-range correlations in the probability distribution. Moreover, although some empirical estimates for the correlation length have previously been given [14, 15], the present tests are the first ones that quantitatively show that the correlation length lies very close to the longer lag parameter p. This indeed means that the errors are due to triple correlations in the GFSR algorithm. Furthermore, for the generators based on a judicious decimation (e.g., $k = 3, 5, 7$) of GFSR generators (with two lags) or when primitive pentanomials are used as a basis for a generator, our results show that similar behavior is observed, but with much weaker correlations. Thus generators using three consecutive exclusive-or operations shuffle bits much better than Rp generators in which only one exclusive-or operation is used. This results from the fact

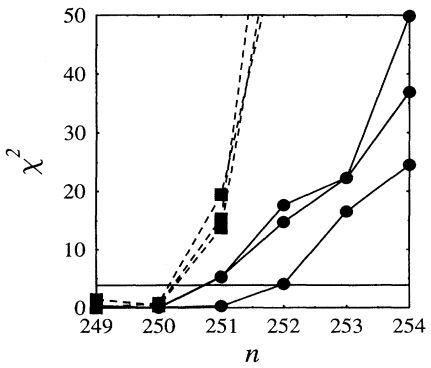

FIG. 5. The χ^2 values for R250 in the n-block test. Curves with circles and squares correspond to $N = 10^8$ and $N = 10^9$ samples, respectively. In both cases three independent runs have been performed. The horizontal line denotes $\chi^2 = 3.841$.

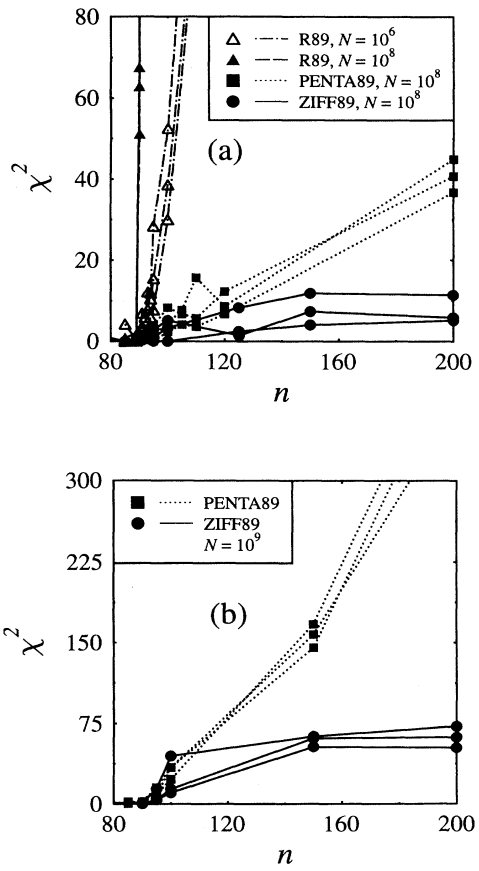

FIG. 6. (a) χ^2 values for R89, PENTA89, and ZIFF89 as a function of the block size n, when $N = 10^8$ samples have been taken. In the case of R89, results with $N = 10^6$ samples (open triangles) are also shown. (b) Results for PENTA89 and ZIFF89 have been compared with each other with better statistics, $N = 10^9$.

that, compared with Rp generators, in ZIFFp generators the three-point correlations are much farther apart, and therefore higher order correlations dominate [22]. Although we are not aware of any theoretical studies for PENTAp generators, we could assume that this is what happens for them also. In other words, the approach of using multiple exclusive-or operations does not actually remove the correlations but makes them weaker.

V. SUMMARY AND DISCUSSION

Modern high-speed computers and demanding applications, such as high-precision Monte Carlo simulations in physical sciences [3] and stochastic optimization [4], have greatly increased the need for fast but reliable random number generation. To this end, efficient tests of random number algorithms and generators are needed. This calls for new empirical and theoretical tests. Although theoretical tests based on studying some properties of algorithms give us basic knowledge of the properties of PRNG's, random number testing remains mainly an empirical science. Though no empirical test can ever *prove* the "goodness" of any random number sequence, such tests give us a valuable insight into their properties. However, since the number of possible tests is practically unlimited, the most important tests should be such that they mimic the properties of the applications in which the random number sequences will be used. This idea naturally leads to the concept of application specific testing introduced in the present work.

We have presented and analyzed four simple tests for detecting correlations in random number sequences. The cluster test is based on the idea of comparing the cluster size distribution of a random lattice with the Ising model at an infinite temperature. Our analysis shows that it is particularly efficient in finding periodic correlations on bit level. Another test based on the use of the Ising model is the autocorrelation test, in which integrated autocorrelation times of some quantities of the Ising model are calculated. This test is very sensitive to correlations in successive random numbers which are used in the cluster formation process of the Wolff algorithm. The two other tests which are based on ideas of random walks, namely the random walk and n-block tests, can be used to quantitatively find the range of correlations for many generators.

As far as the PRNG's are concerned, our analysis shows that the origin of the errors observed in Refs. [12–16] for GFSR generators must be due to the triple correlations. We have also tested a set of generators which should be able to avoid this problem by using four lags instead of two. Such generators can be formed by using tables of primitive pentanomials [29] or by decimating GFSR sequences [22,37]; i.e., taking every third number of their sequence, for example. Our results show that such approaches do not completely eliminate correlations but do make them much weaker and therefore greatly improve the quality of generated random numbers. Such generators indeed passed all our tests, when the longest lag parameter p was chosen large enough ($p \geq 1279$).

Finally, we would like to point out that our results are not only restricted to those particular models which have been studied in this work but have relevance in other applications as well. For example, in connection with studies of other models using random walks such as percolation phenomena and diffusion limited aggregation, those generators which failed our tests should be avoided.

In conclusion, our aim has been to introduce a set of tests which are application specific from the point of view of Monte Carlo computer simulations in particular. Practical details of the algorithmic implementation of an actual test bench based on these tests will be published separately [19]. We hope that the present tests can be used to design better generators for demanding applications in physical sciences, and that further work will be done on developing suitable tests for other applications also.

ACKNOWLEDGMENTS

I. V. wants to thank R. M. Ziff for correspondence and giving his results prior to publication, D. Stauffer for correspondence, and the Jenny and Antti Wihuri Foundation, Finnish Academy of Sciences, and Neste Foundation for financial support. This work has also been supported by the Academy of Finland. Computational resources at University of Helsinki and Tampere University of Technology are gratefully acknowledged.

[1] N. Metropolis and S. Ulam, J. Am. Stat. Assoc. **44**, 335 (1949).
[2] N. Metropolis, A. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[3] K. Binder, in *Monte Carlo Methods in Condensed Matter Physics*, edited by K. Binder (Springer-Verlag, Berlin, 1992).
[4] E. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines, A Stochastic Approach to Combinatorial Optimization and Neural Computing* (John Wiley & Sons, Chichester, 1989).
[5] T. E. Hull and A. R. Dobell, SIAM Rev. **4**, 230 (1962).
[6] D. E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, 2nd ed. (Addison-Wesley, Reading, MA, 1981), Vol. 2.
[7] P. L'Ecuyer, Ann. Oper. Res. **53**, 77 (1994).
[8] I. Vattulainen, `cond-mat@babbage.sissa.it` No. **9411062** (1994).
[9] G. J. Chaitin, J. Assoc. Comput. Mach. **13**, 547 (1966); Sci. Am. **232**, 47 (1975); J. Assoc. Comput. Mach. **22**, 329 (1975); Sci. Am. **259**, 52 (1988).
[10] A. Compagner, Am. J. Phys. **59**, 700 (1991).
[11] P. Martin-Löf, Inf. Control **9**, 602 (1966).
[12] P. D. Coddington, Int. J. Mod. Phys. C **5**, 547 (1994).

[13] A. M. Ferrenberg, D. P. Landau, and Y. J. Wong, Phys. Rev. Lett. **69**, 3382 (1992).
[14] P. Grassberger, J. Phys. A **26**, 2769 (1993).
[15] P. Grassberger, Phys. Lett. A **181**, 43 (1993).
[16] W. Selke, A. L. Talapov, and L. N. Shchur, Pis'ma Zh. Eksp. Teor. Fiz. **58**, 684 (1993) [JETP Lett. **58**, 665 (1993)].
[17] K. Kankaala, T. Ala-Nissila, and I. Vattulainen, Phys. Rev. E **48**, 4211 (1993).
[18] I. Vattulainen, T. Ala-Nissila, and K. Kankaala, Phys. Rev. Lett. **73**, 2513 (1994).
[19] I. Vattulainen and T. Ala-Nissila, Comput. Phys. (to be published).
[20] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, London, 1982).
[21] U. Wolff, Phys. Rev. Lett. **62**, 361 (1989).
[22] R. Ziff (unpublished).
[23] S. L. Anderson, SIAM Rev. **32**, 221 (1990).
[24] F. James, Comput. Phys. Commun. **60**, 329 (1990).
[25] P. L'Ecuyer, Comm. ACM **33**, 86 (1990).
[26] T. G. Lewis and W. H. Payne, J. Assoc. Comput. Mach. **20**, 456 (1973).
[27] J. R. Heringa, H. W. J. Blöte, and A. Compagner, Int. J. Mod. Phys. C **3**, 561 (1992).
[28] S. Kirkpatrick and E. P. Stoll, J. Comput. Phys. **40**, 517 (1981).
[29] Y. Kurita and M. Matsumoto, Math. Comput. **56**, 817 (1991).
[30] N. Zierler, Inf. Control **15**, 67 (1969).
[31] N. Zierler and J. Brillhart, Inf. Control **13**, 541 (1968).
[32] S. K. Park and K. W. Miller, Comm. ACM **31**, 1192 (1988).
[33] *Convex Fortran Guide*, 1st ed. (Convex Computer Corp., Richardson, 1991), p. 553.
[34] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes, The Art of Scientific Computing*, FORTRAN version (Cambridge University Press, Cambridge, England, 1989), p. 199.
[35] G. Marsaglia and A. Zaman, Stat. Prob. Lett. **8**, 35 (1990).
[36] M. Fushimi, Appl. Math. Lett. **2**, 135 (1989).
[37] R. M. Ziff, Phys. Rev. Lett. **69**, 2670 (1992).
[38] K. Pearson, Philos. Mag. **50**, 157 (1900).
[39] M. F. Sykes and M. Glen, J. Phys. A **9**, 87 (1976).
[40] D. Stauffer, *Introduction to Percolation Theory* (Taylor & Francis, London, 1985).
[41] J.-S. Wang and R. H. Swendsen, Physica A **167**, 565 (1990).
[42] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer-Verlag, Berlin, 1988).
[43] N. S. Altman, SIAM J. Sci. Stat. Comput. **9**, 941 (1988).
[44] I. Vattulainen, K. Kankaala, J. Saarinen, and T. Ala-Nissila, Comput. Phys. Commun. **86**, 209 (1995).
[45] U. Wolff, Phys. Lett. B **228**, 379 (1989).
[46] Using Eq. (4.13) in a reference [A. E. Ferdinand and M. E. Fisher, Phys. Rev. **185**, 832 (1969)], we found $E \approx 1.453\,12$. Recently, J. Tobochnik has informed us that a more precise value may be $E \approx 1.453\,06$. Although this difference is not relevant in the present work, there are certainly applications in which the exact value for E must be determined more accurately than in this study.
[47] We note that in this work we have considered only the case where the empirical distribution is too far from the theoretical one. In other words, a generator fails the test if the χ^2 value is "too large." However, one can also take into account the case where the empirical distribution follows the expected one too smoothly; i.e., the χ^2 value is "too small."
[48] Since the modulus of RAND is a power of 2, its least significant bit has a period of 2, the second least significant a period of 4, and so on. For more details, see W. F. Eddy, J. Comput. Appl. Math. **31**, 63 (1990).
[49] S. W. Golomb, *Shift Register Sequences*, revised edition (Aegean Park Press, Laguna Hills, 1982), pp. 78–79.
[50] A. Compagner and A. Hoogland, J. Comput. Phys. **71**, 391 (1987).