Pattern Count Statistics for the Analysis of Time Series in Mind-Matter Studies

WERNER EHM

Institut für Grenzgebiete der Psychologie und Psychohygiene
Wilhelmstr. 3a, D-79098 Freiburg, Germany

Abstract—A class of statistical tests is proposed for testing the null hypothesis that a given time series is random in the sense of pattern-free. The basic idea is to look whether certain patterns occur more frequently in the time series than expected under the null hypothesis. The construction and theoretical basis of the tests are an application of martingale theory. The method is applied to data from an experiment designed to study possible correlations between the intention of human operators and a physical random number generator. In order to account for possible heterogeneity of the deviations from the null hypothesis, a new combination technique is introduced for the aggregation of individual results across agents and/or intentions. Significant results are found with respect to oscillatory patterns and intention "baseline", but are due to one single operator, essentially. Some methodological issues related to multiple testing are addressed.

Keywords: (testing for) mind-matter interrelations—random number generator—martingale—pattern analysis—order pattern—local variation pattern—combination of p-values

1. Introduction

During the past decades the possibility of mind-matter interrelations has been explored in a large number of experiments. According to one basic paradigm the material domain is represented by a physical device which produces random bits (sequences of zeros and ones), while the mental domain is operationally represented by an intention of a human operator (subsequently called "agent"); cf. Jahn et al. (1997), and references given therein. The intention usually is prescribed by the experimental protocol. In a successful experiment it should be correlated with output of the physical device deviating significantly from its standard (random) output, expected if the device operates in its null mode; that is, if it is not exposed to conditions possibly affecting its "autonomous" random behavior.

To get a sound basis for the statistical analysis of such experiments, the standard output of the physical device has to be gauged in order to give rise to a viable null hypothesis. In the original implementation of the experiments at the Princeton Engineering Anomalies Research (PEAR) laboratories (Jahn et al., 1997, 2000) the primary bits are combined and processed in such a way that the (null mode) final output x_t, $t = 1, 2, \ldots,$ of the "random number generator" (RNG) can be regarded as a sample from the binomial distribution $\mathcal{B}(N, p)$ with parameters $N = 200$ (number of Bernoulli trials) and $p = 1/2$ (success probability). Therefore, *under the null hypothesis* $\mathcal{H}_0$ *the time series* $\{x_t,\ t = 1, \ldots, \tau\}$ *produced by the RNG is supposed to be the realization of an i.i.d. sequence of random variables with (common) distribution* Q [where $Q = \mathcal{B}(200,\ 1/2)$ in the case of the PEAR implementation]. The abbreviation "i.i.d." means "independent and identically distributed". Extensive tests with calibration data produced by the PEAR RNG showed no significant deviations from this null-hypothesis.

For testing $\mathcal{H}_0$ some idea is required as to what kinds of deviations from the i.i.d. Q behavior might occur under the experimental conditions. Traditionally, a shift of the mean value of Q is, explicitly or implicitly, assumed as the alternative hypothesis. For instance, if $Q = \mathcal{B}(N,\ p_0)$ under $\mathcal{H}_{0,}$ then under the alternative the random numbers x_t are assumed to be i.i.d. with distribution $\mathcal{B}(N,\ p)$, $p \neq p_0$, where the direction of the hypothesized deviation of p from p_0 conforms to the experimental condition. More recently, deviations from $\mathcal{H}_0$ have been studied which, unlike in the mean shift models, allow for statistical dependence between the random numbers x_t. Related proposals in the literature include analyses based on scaling indices (Atmanspacher et al., 1999, 2001), Hurst coefficients (Pallikari and Boller, 1999), and harmonic analysis (Dobyns, 1999), among others.

In this paper we introduce a class of tests based on the frequency of the occurence of specific patterns in a time series. The envisaged deviations from $\mathcal{H}_0$ are similar to those detectable by scaling indices. The idea is that from time to time some pattern occurs in the random process which would be unlikely to occur under $\mathcal{H}_0$. Our tests are tailored to detect precisely defined patterns with high power. They thus complement scaling index-based tests, which are omnibus-type tests, i.e., tests that can detect a large variety of deviations from $\mathcal{H}_0$, though often with low power; cf. Atmanspacher et al. (2001). Out of the multitude of possible patterns a certain range will be selected by introducing a few pattern types and systematically varying the free parameters. The related tests permit an individual statistical assessment of the data set associated with each agent/intention pair. For combining the individual results across agents and/or intentions we propose a new kind of summary statistic which does not presuppose homogeneity of the possible deviations from $\mathcal{H}_0$ across agents. It has improved detection power in situations where such deviations occur in a few cases only.

Any statistical analysis aiming at detailed information, i.e., trying to answer *several* questions, raises a multiple testing problem. Basically, the approach of the present paper calls for a compromise between two conflicting desires: to perform a large number of different tests in order to get a broad knowledge of which patterns are conspicuous and which are not, and to "know" with statistical confidence by controlling the error probability of the first kind that "significant" outcomes are due to mere chance fluctuations. Our strategy is to maintain statistical confidence *partially*: construct (joint) tests which properly account for the relevant stages of data aggregation up to a certain level, separately for each pattern; beyond that, make suggestions how one could integrate the individually rigorous results into a broader picture, and admit the exploratory character of any related conclusion. In principle, multiple statements with confidence, referring to several patterns, could be made by constructing yet more inclusive combination tests or by simple Bonferroni adjustments to the p-values. However, due to complex correlations which exist between the single components of a combination procedure at such high levels, construction of valid joint tests becomes definitely non-trivial, and Bonferroni corrections tend to be too conservative.

Section 2 of this paper introduces the pattern count approach and proposes concrete classes of patterns defined by order relations and by local variation characteristics. The results are presented in section 3, along with the procedure used to pool the individual outcomes across agents and/or intentions. A summary of our approach and the major results concludes the main part of the paper in section 4. An appendix provides mathematical details about the pattern count tests and develops their distribution theory using martingale methods.

2. Data, Patterns, and Tests

The data to be analysed were collected by the Freiburg Anomalous Mind/Machine Interactions group (FAMMI) as a part of the recent experimental replication studies of the MMI consortium. For detailed information about these experiments see Jahn et al. (2001).

Each single data set consists of a series of random bit counts x_t of total length $\tau = 10000$. Under the null-hypothesis $\mathcal{H}_0$ the x_t are considered as i.i.d. $\mathcal{B}(200, 1/2)$-distributed random variables. As a feedback to the agent, the centered successive partial sums $\sum_{s \leqslant t}(x_s - 100)$ are presented graphically on a computer screen. For every agent there are three time series corresponding to different conditions: under conditions HI(GH) or LO(W) the agent's task is to "make the partial sum curve go up or down", respectively, while under condition BL (BASELINE) no specific task is being posed. The data to be analysed come from 16 participants of the first phase of the FAMMI experiments. The possible deviations from $\mathcal{H}_0$ envisaged here are related to the occurence of two types of patterns: order patterns and local variation patterns.

The speculative guiding idea suggesting these patterns is that any process responsible for correlations between agent intention (or: related brain processes) and the physical device should be very subtle, making its build-up and maintenance prone to the faintest disturbances. Such a process may be supposed to get started only sporadically and "stay active" for short time intervals only (whatever "short" may mean in the absence of knowledge about relevant time scales). One thus arrives at the subsequently defined patterns in a fairly natural way, although other implementations are conceivable as well.

For ease of presentation a minimum of technicalities is required. Quite generally, we consider any subset B of the L-dimensional Euclidean space R^L as a "pattern of length L". (Any point of R^L may be described by its coordinates, i.e., by L real numbers.) Given a time series $x_1, x_2, \ldots$ and a pattern B of length $L \geqslant 1$ we say that "the pattern B occurs at time t" if the "delay vector" $(x_{t-L+1}, \ldots, x_t)$ consisting of the L data points corresponding to time points $t-L+1, t-L+2, \ldots$ up to t falls into the set B. The basic statistics is obtained by counting how often pattern B occurs in the given time series.

2.1 Order Patterns

These patterns are defined in terms of order relations. For example, "*increasing triplets*" correspond to the pattern $B = \{y = (y_1, y_1, y_3) \in R^3 : y_1 < y_2 < y_3\}$. It occurs at time t if $x_{t-2} < x_{t-1} < x_t$. "*Alternating quadruples*" are described by the set union $B = B_u \cup B_d$, where the "alternating-up" and "alternating-down" patterns $B_u = \{y \in R^4 : y_1 < y_2, y_2 > y_3, y_3 < y_4\}$ and $B_d = \{y \in R^4 : y_1 > y_2, y_2 < y_3, y_3 > y_4\}$ end with an upward or downward step, respectively. Increasing/decreasing and alternating (up/down) patterns of arbitrary length $l(\geqslant 2)$ are defined analogously. The respective pattern classes are indicated by *inc, dec, altu, altd.*

A different type of order pattern, also referred to as a *tube pattern,* is the "*L-in*-[c_1, c_2]" pattern $B = [c_1, c_2]^L$, where $c_1 \leqslant c_2$ and $L \geqslant 1$ are fixed parameters. It occurs at time t if $c_1 \leqslant x_{t-k} \leqslant c_2$ for $k = 0, \ldots, L-1$.

Examples. The series $x = \{96, 91, 102, 99, 103, 115, 93, 98, 81, 95, 102, 98\}$ of *length* $\tau = 12$ contains two increasing triplets occurring (i.e., ending) at "time" $t = 6$, 11; no decreasing triplet; two 3-*in*-[0, 99] patterns occurring at $t = 9$, 10; and 5 alternating quadruples, with *altu* and *altd* occurring at $t = 5$, 8, 10, and at $t = 4$, 9, respectively.

2.2 Local Variation Patterns

These patterns capture quantitative aspects of the local fluctuations of the x_t's. Their definition involves a local variation statistic measuring the Euclidean distance D_t of the delay vector $(x_{t-d+1}, \ldots, x_t)$ from a vector of reference coordinates. We consider three versions, with distances D_t defined by

$$D_t^2 = \sum_{i=0}^{d-1} (x_{t-i} - 100)^2; \tag{LV1}$$

$$D_t^2 = \sum_{i=0}^{d-1} (x_{t-i} - x_{t-d})^2; \tag{LV2}$$

$$D_t^2 = \sum_{i=0}^{d-1} (x_{t-i} - x_{t-i-1})^2. \tag{LV3}$$

D_t^2 in (LV1) measures the mean square distance from the common mean value 100 and thus represents a running χ^2 statistic. Equations (LV2) and (LV3) are variants of (LV1) with data-dependent reference points, namely the "d before t"

and the "immediately preceding" data points, respectively. A local variation pattern is defined by setting a limit to the magnitude of D_t: the pattern occurs at time t if $D_t \leqslant \rho$ (for some constant ρ). Note that the event ($D_t \leqslant \rho$) can be written in the form ($y_t \in B$) for some $B \subset R^L$, where for (LV1) the appropriate delay vectors are $y_t = (x_{t-d+1}, \ldots, x_t)$ (length $L = d$), whereas for (LV2) and (LV3) one has to take $y_t = (x_{t-d}, \ldots, x_t)$ (length $L = d + 1$). We actually will consider two-sided local variation patterns obtained by setting upper *and* lower limits, i.e., by counting how often D_t is $\leqslant \rho_1$ or $\geqslant \rho_2$, where $\rho_1 < \rho_2$. (Of course, this is equivalent to counting how often D_t lies *between* ρ_1 and ρ_2.) They will be referred to as *LV*1, *LV*2, and *LV*3 patterns, respectively, depending on the underlying distance function.

2.3 Fixing the Pattern Parameters

The restriction to the above pattern classes still leaves considerable freedom in the choice of the pattern parameters such as, e.g., the length of an alternating pattern. Because factually justifiable criteria for special parameter values are missing, we will always consider a *grid* of such values, the grid itself being chosen by criteria derived from theoretical statistical considerations. Thus, parameter choices are not entirely ad hoc and do *not* depend on the experimentally observed data.

In case of *order* patterns we "grid" the pattern length L, i.e., we consider all patterns (of a certain type) of length $L \leqslant L_{\max}$. If $L_{\max}$ is too large, the corresponding pattern occurs too rarely for "sufficiently good statistics". A relevant quantity in that respect is the minimal eigenvalue $\lambda_{\min}$ of the predictable covariation matrix V_τ, which should be "large"; cp. Appendix A.2. The rule adopted here is to choose $L_{\max}$ maximally subject to the condition that $\lambda_{\min}$ (computed from the average of simulated V_τs) be not smaller than 20 (approximately). This rule, along with an essentially arbitrary choice of intervals for the tube patterns, leads to the values displayed in Table 1.

By the symmetry of the binomial distribution $\mathcal{B}(200, 1/2)$, the same values, $L_{\max} = 5, \lambda_{\min} = 30.0$, apply also to the *dec* patterns. Likewise, *altu* and *in* [105, 200] may be replaced by their mirror patterns *altd* and *in* [0, 95].

The expected absolute frequency of the patterns of length $L_{\max}$ (in a binomial, length $\tau = 10000$ time series) is roughly between 50 and 80, with one exception: for *altu/d* patterns of length 8 the average pattern frequency is almost 300. Therefore, we depart from the $\lambda_{\min}$ rule in this case and consider *altu/d* patterns up to length $L_{\max} = 11$, which still occur roughly 70 times on the average. Due to the high correlation between the counts for patterns of neighboring lengths, $\lambda_{\min}$ then drops to 4.3, but probability plots of simulated z-scores indicate that the normal approximation remains quite accurate.

The parameters for the *local variation* patterns are fixed as follows. The pattern length is taken as $L = 4$ in case of the *LV*2 and *LV*3 patterns, and as $L = 3$ for *LV*1 (so that the variables D_t^2 have the same number $d = 3$ of summands

TABLE 1
Maximal lengths of order patterns.

Pattern type	*altu*	*inc*	*in* [97, 103]	*in* [105, 200]
$L_{\max}$	8	5	5	4
$\lambda_{\min}$	18.1	30.0	24.9	23.6

in each case; see the above definitions). This particular choice is made for the sake of compatibility with the investigations using scaling index analysis (cf. Atmanspacher et al., 1999).

To motivate the choice of the grid of radii we first note that for *LV*1 the random variable $D_t^2/50$ $[50 = \mathrm{Var}(x_t)]$ is approximately χ^2 distributed with 3 degrees of freedom (under $\mathcal{H}_0$). For *LV*2 and *LV*3, $D_t^2/50$ also has three summands, and a twice as large expectation, $\varepsilon = 6$. In these cases a χ_3^2 approximation to the distribution of $D_t^2/100$ is not justified even asymptotically, due to dependency among the sum members. Nevertheless, it yields a crude guideline for the choice of the radii. Considering that values in the tails of the χ^2 distribution are of primary interest, we end up with the following choice: Let $\gamma(\alpha)$ and $\delta(\alpha)$ denote the lower and the upper α-quantile of the χ^2 distribution with 3 degrees of freedom, respectively [so that a χ_3^2-distributed random variable is $<\gamma(\alpha)$ or $>\delta(\alpha)$ with probability α each]. Let α run through the grid $\alpha_k = k/100, 1 \leqslant k \leqslant 10$ $(0.01, 0.02, \ldots, 0.10)$. Then in the one-sided case we set $\rho_k = \sqrt{\varepsilon * 50 * \gamma(\alpha_k)}$, with $\varepsilon = 3$ for *LV*1 patterns, and $\varepsilon = 6$ for *LV*2 and *LV*3 patterns. In the two-sided case the lower limits coincide with the one-sided values, $\rho_{k,1} = \rho_k$, and the upper limits are $\rho_{k,2} = \sqrt{\varepsilon * 50 * \delta(\alpha_k)}$. Table 2 provides a (partial) list of the numerical values.

Multiplying these ρs by $\sqrt{2}$ gives the radii for the *LV*2 and *LV*3 patterns.

2.4 Interpretation of Patterns

"Interpretation" here does not allude to any concrete model of possible mind-matter interrelations. Rather, it is to be understood as a statement that a pattern conforms to, or reflects somehow, the phenomena one would be prepared to observe under a specific experimental condition.

For example, *in* $[c_1, c_2]$ patterns, where $c_1 > 100$, reflect that the feedback curve (centered partial sums) turns upward (locally) in accordance with the behavior one would be prepared to see under condition HI. Moreover, if those

TABLE 2
Radii for *LV*1 patterns.

α_k	0.01	0.02	0.03	0.04	0.05	...	0.10
$\rho_{k,1}$	2.3962	3.0400	3.5007	3.8740	4.1943	...	5.4054
$\rho_{k,2}$	23.8169	22.1782	21.1510	20.3853	19.7671	...	17.6796

patterns occur more frequently than expected under $\mathcal{H}_0$, then there will be a tendency that the feedback curve's endpoint goes up, as supposed in the traditional (mean-value oriented) paradigm.

Monotone (in- or decreasing) segments intuitively may be associated with conditions HI/LO, in the sense that they may be seen as representing sporadic "success" in regard to the respective task. However, the up and down of the feedback curve is no unique indicator for such patterns: if all values of an *increasing* segment lie *below* the mean value 100, the feedback curve goes down. Moreover, the statistical effect of an increased occurrence of such patterns on the feedback curve's endpoint is less clear than in the case considered above. Related pattern analyses therefore appeal to a more general class of intentional correlates than in the traditional approach.

The association of alternating segments with the experimental condition BL is not immediate if the latter is seen as a pure control condition. It appears as more natural, however, if BL is understood as the task to keep the feedback curve close to the baseline, as some agents may have done. Alternating segments are but one way of describing "temporal (sporadic, local) success" in attempts to comply with this task. Another, quantitative one is provided by the local variation patterns: weakly fluctuating segments of the time series yield low values of the local variation measures, so an increased frequency of low values indicates sporadic "success".

2.5 Statistical Analysis

A detailed description of the procedures used in the statistical analysis is given in the Appendix. Here is a brief account of the basic notions and steps.

For each single pattern the number of its occurences is at first transformed into a z-score-type statistic (a random variable which, ideally, is standard normally distributed under the null hypothesis $\mathcal{H}_0$). For technical reasons, the necessary standardization differs from the usual one and depends on the theory of martingales; see Appendix A.1. Tests of $\mathcal{H}_0$ are constructed by combining the z-scores of several single patterns belonging to a certain type into one test statistic. For example, when dealing with patterns of the type *inc* there are four z-scores for the four single *inc* patterns of (fixed) lengths $2, \ldots, 5$. Among the many recipes of "cooking" a test statistic from the single pattern z-scores, we focus on two well-established ones, namely the χ^2 type and maximum (M) type tests. Due to statistical dependence of the single pattern z-scores, the χ^2 test reduces to a familiar sum of squares only after a suitable multivariate renormalization. Conversely, the test statistic for the M test is defined directly in terms of the "raw" z-scores, namely as the maximum of their absolute values. One-sided versions of this test are required if more differentiated information is desired: does a pattern occur *too often* or *too rarely* compared with $\mathcal{H}_0$? The corresponding tests are referred to as M_+ and M_- tests, respectively; see Appendix A.2.

3. Application to RNG Data: Results

In order to structure the material, the experimental and the statistical side of the analysis are distinguished. At the experimental side there are 48 experimental units, or "cases", consisting of the 16 (agents) times 3 (intentions) agent/intention pairs and the associated time series. At the statistical side there are four tests each for every pattern, namely the χ^2, M, M_+, and M_- tests. Results are obtained on different levels. On the lowest, or *individual* level, one associates with each case (i.e., agent/intention pair) the outcomes of the four tests applied to each pattern. The outcomes are reported in terms of p-values. Initially, there will be *two* approximate p-values for every case-pattern-test combination, determined by the different methods described in Appendix A.2.[1] Replacement of those values by one single p-value will constitute the only kind of aggregation at the statistical side.

At the experimental side there will be two stages of aggregation. In the first stage the individual results (the p-values obtained on the individual level) are combined across agents and distinguished by intention. In the second aggregation step, leading to the *global* level, results are combined also across intentions.[2]

It is important to clarify the meaning attached to p-values within the present framework. Generally, all p-values have their usual meaning for the special situation they are calculated for. That always includes *partial* adjustment for multiple testing. For example, each single p-value among those plotted in the figures below accounts for the fact that the related test is a function of the *smallest* p-values among 48 (or 16, respectively) cases. On the other hand, statistical assertions (claiming "significance") reaching beyond the special situation under consideration cannot be made, and any broader statement has to be considered as tentative and exploratory. This specifically applies to statements involving several pattern/test pairs simultaneously.

As for the results proper, we begin with a case study on the individual level. Special attention is payed to the case 16/BL—this notation is short for "agent #16, intention BL"—for which noticeable deviations from randomness were found earlier (Atmanspacher et al., 1999). In section 3.2 we then introduce a method of combining the single p-values of the individual agents and discuss the ensuing results on the global level and by intention.

3.1 The Case 16/BL

The p-values for the various patterns and tests are displayed in Table 3. The two entries in each cell are the p-values computed by two different approximation methods: the lower entry is the p-value obtained by method *fullsim,* the upper by method *hybsim* for the M-type tests, or from the asymptotic χ^2 distribution for the χ^2 test; cp. Appendix A.2.

The table contains a number of small p-values identifying the patterns with unexpected frequency of occurrence. Most conspicuous is the *altu* pattern, where

TABLE 3
Individual p-values for case 16/BL. Lower entries: p-values obtained by *fullsim.*
Upper entries: p-values obtained by *hybsim* (M-type tests),
or from the asymptotic chi-square distribution (χ^2 test).

	χ^2	M	M_-	M_+
altu	0.00025	0.00012	0.99882	0.00005
	0.00018	0.00018	0.99850	0.00006
altd	0.05694	0.44649	0.22828	0.60582
	0.05694	0.44484	0.23456	0.60058
inc	0.02534	0.39580	0.20264	0.26750
	0.02530	0.39450	0.20300	0.26336
dec	0.00553	0.00481	0.00245	0.80895
	0.00532	0.00496	0.00204	0.80506
in [105, 200]	0.12198	0.03692	0.01821	0.94637
	0.12222	0.03666	0.01680	0.94400
in [0, 95]	0.62534	0.66305	0.35845	0.81286
	0.62626	0.66478	0.36430	0.80660
in [97, 103]	0.76038	0.40614	0.92938	0.20841
	0.75958	0.40822	0.92892	0.20336
*LV*1	0.00077	0.00586	0.99746	0.00281
	0.00092	0.00574	0.99804	0.00320
*LV*2	0.74808	0.42957	0.93346	0.21963
	0.74442	0.42760	0.93680	0.21642
*LV*3	0.05180	0.01473	0.94009	0.00706
	0.05216	0.01464	0.94360	0.00736

the individual z-scores for pattern lengths 3 to 11 are persistently high (namely 2.86, 4.01, 3.42, 3.23, 4.00, 4.39, 4.35, 2.74, 2.87), yielding a p-value of about $6 *10^{-5}$ for the M_+ test.

As it seems, deviations mainly are connected with the local fluctuations of the x_ts, which are either too strong or too weak (with respect to $\mathcal{H}_0$); cp. the M_+ test p-values for the patterns *altu*, *LV*1, *LV*3. The small figures in Table 3 indicate that the corresponding patterns occur too frequently. Note also that the trend-like patterns *dec, in* [105, 200] occur too rarely; cp. the M_- test column. The χ^2 test is less sensitive than the M type tests against large deviations in one component of the z-score vector; it rather "collects" smaller, unsystematic inconsistencies. Therefore, incoherent results of the M and χ^2 tests as for pattern *inc* are not surprising. On the side of the patterns, too, there appear to be inconsistencies. For instance, both the *LV*1 and the *in* [97, 103] pattern measure "closeness to the mean 100", in a sense. Yet, the test results show little concordance. Similarly, patterns *altu, LV*1, *LV*3 stand out, but the related patterns *altd, LV*2 do not. Such inconsistencies may result from the connection between the set theoretic properties of the patterns and the correlations between the corresponding count statistics. For example, *altu* and *altd* patterns are *disjoint* as point sets, and the related count statistics are therefore *negatively* correlated; cp. the corresponding remark near the end of Appendix A.1.

A comparison of the p-values in each cell of Table 3 shows good agreement between the two methods used to calculate approximate p-values. The largest relative error, 0.5, occurred for pattern *altu* and the *M* test, where the *p*-values are 0.00012 and 0.00018, respectively. (The relative error between two *p*-values p_1, p_2 is computed as $|p_1 - p_2|/\mu$, where μ equals the minimum of $\min\{p_1, p_2\}$ and $1 - \max\{p_1, p_2\}$.) A more systematic assessment of these findings, including a discussion of the Monte Carlo sampling variability, may be found in Appendix A3.

In interpreting these *p*-values it has to be observed that case 16/BL received special consideration because of the results of the scaling index analysis of Atmanspacher et al. (1999). It is not straightforward to adjust the *p*-values properly for this selection. However, a simple, conservative estimate can be obtained by pretending that case 16/BL was selected because of its having, separately for every pattern/test pair, the smallest *p*-value among all 48 cases (or 16 cases, respectively, when treating each intention on its own). Under that surmise a test would be (individually) significant at the 5% level if the *p*-value was below 0.00107 (selection among 48 cases) or 0.0032 (16 cases), respectively. Re-inspecting Table 3 from this angle shows that four among the 40 pattern/test pairs pass the more restrictive of the two criteria. Of course, this estimate depends on an incorrect assumption: although case 16/BL happens to have the minimal *p*-value for some pattern/test pairs, it generally has not. The next subsection proposes a more adequate approach where all cases are treated on an equal footing.

3.2 Combination of Small p-Values: Theory

The basic assumption underlying our approach is that deviations from the null hypothesis, if any, are heterogeneous across the cases and confined to a few at most. Under such circumstances most cases contribute "noise", while for a few of them the deviation from $\mathcal{H}_0$ manifests itself by a small *p*-value. This suggests to disregard most of the individual *p*-values and to construct a summary test depending on the smallest *p*-values only. Of course, proper adjustment has to made for this selection procedure.

Beforehand let us recall the following facts: if a test statistic *T* is distributed with the continuous cdf (cumulative distribution function) $F_0(t)$ under $\mathcal{H}_0$, then the random variable $p = 1 - F_0(T)$ has the uniform distribution $\mathcal{U}$ on [0, 1]. Conditionally on the value of *T*, the "*p*-value" *p* equals the (conditional) probability that a variable T^* with cdf $F_0(t)$ (i.e., distributed like *T*) exceeds the observed value *T*. If we have *N* *p*-values from independent experiments then the *p*-values arranged in increasing order, $p_{(1)} < p_{(2)} < *** < p_{(N)}$, have the same joint distribution under $\mathcal{H}_0$ as the order statistics of a sample of size *N* from $\mathcal{U}$. ($\mathcal{H}_0$ here denotes the amplified null hypothesis saying that the [former] $\mathcal{H}_0$ holds in every single experiment.)

Our *p*-value combination criterion *CC* is defined as the sum of the smallest *m* *p*-values, $CC = p_{(1)} + *** + p_{(m)}$, where *m* is a positive integer yet to be fixed. Small values of *CC* speak against the null hypothesis. The rationale behind the criterion is as follows. Suppose that deviations from $\mathcal{H}_0$ actually occur in m_0

cases. Then compared with $\mathcal{H}_0$ (which would correspond to $m_0 = 0$), the distributions of the corresponding p-values are shifted to the left (to smaller values) in some sense. The same should then hold for the distribution of CC, no matter whether m is less or larger than the number m_0 of deviant cases (provided only that $m_0 \geqslant 1$).

The null distribution of CC can be determined explicitly. Suppose that the p-values $p_1, \ldots, p_N$ are distributed like an i.i.d. sample from $\mathcal{U}$. Then the cdf $Q(x)$ $(0 \leqslant x \leqslant 1)$ of CC is given by

$$Q(x) = \sum_{k=1}^{m} (-1)^{m-k} \frac{k^m}{k!(m-k)!} \Big] \, 1 - \Big[\max\Big\{1 - \frac{x}{k}, 0\Big\}\Big]\Big]^N \Big[\qquad (1 \leqslant m \leqslant N).$$

This can be proven for instance by conditioning on $p_{(m)}$ and noting that the sum of the other terms in CC divided by $p_{(m)}$ equals a sum of i.i.d. random variables, the distribution of which is known. Alternatively, one may use a suitable representation of order statistics along with calculations involving Fourier transforms and the residue calculus of complex analysis.

For consistency of presentation the criterion reported below will actually be the p-value associated with CC, namely the *transformed* quantity $C = Q(CC)$. It has the uniform distribution $\mathcal{U}$ under the null hypothesis $\mathcal{H}_0$, with small values of C speaking against $\mathcal{H}_0$.

When applying the combination criterion one is faced with the problem that the p-values stemming from the individual cases are, though independent, *not exactly* uniformly distributed under $\mathcal{H}_0$. This is because for none of the four test statistics (applied to any of the patterns) the exact null-cdf F_0 is known. Available are only the approximations or estimates $\hat{F}_0$ determined by the methods described in Appendix A.2. Thus we only have approximate p-values $\hat{p} = 1 - \hat{F}_0(T)$ (T stands for any of the four test statistics), which are only approximately $\mathcal{U}$-distributed under $\mathcal{H}_0$. Often the consequences are negligible. Here, however, the criterion depends on extreme (small) p-values, where approximations tend to become inaccurate and statistical procedures based thereupon may become unreliable. A detailed assessment of this important technical problem is given in Appendix A.3. Our ensuing practical conclusions for the present context are as follows: (i) p-values of order smaller than 10^{-3} are little reliable; and (ii) p-values of order larger than 10^{-3} are not substantially biased and fairly reliable. Therefore, criterion C should be sufficiently valid if C does not depend too much on the most extreme p-values. This means that m should not be chosen too small. We tentatively consider $m \geqslant 3$ as sufficient.

3.3 Combination of Small p*-Values: Results*

So far criterion C still is multiply valued: the precise value depends on the approximation method used to determine the single-case p-values. From now on we make C uniquely valued by always taking the *maximum* of the C values resulting from the different approximation methods. Other definitions may be more efficient; however, the present one has the advantage of being conservative.

Good arguments for a particular choice of m are missing (apart from the technical limitations mentioned above). Therefore m will be varied between $m = 1$ and $m = 7$, with the proviso that results for small m, e.g., $m < 3$, should be interpreted with caution. The central range $3 \leqslant m \leqslant 5$ is tentatively considered as the one of major interest.

The large number of (C-related) p-values corresponding to the various patterns, tests, and choices of m calls for a systematicized form of presentation and survey. Thus firstly, results are presented *graphically* by plotting $C = C_m$ against m. (We sometimes add an index m to emphasize the dependence of C on m.) For each pattern there is one such plot containing the results of all four tests. Recall that C is approximately $\mathcal{U}$-distributed under $\mathcal{H}_0$, with small values speaking against $\mathcal{H}_0$. An auxiliary line at height 0.05 is added for ease of orientation.

Secondly, we adopt the following *exploratory* procedure to identify "conspicuous" patterns: every pattern for which at least one of the four tests satisfies the condition $C_m \leqslant 0.05$ for *every* $m = 3, 4, 5$ is marked by the symbol $\langle\ \rangle$. For ease of reference we call any pattern satisfying this condition a "$\langle\ \rangle$ pattern", and any pattern/test pair satisfying $C_m \leqslant 0.05$ for $m = 3, 4, 5$ a "$\langle\ \rangle$ pair". The probability under $\mathcal{H}_0$ that C_m is less than 0.05 for $m = 3, 4, 5$ (simultaneously) is about 0.03, so the condition certainly is not too restrictive.

Figure 1 displays the results for the global level, where the cases consist of all 48 agent/intention pairs. There is only one $\langle\ \rangle$ pattern, $LV1$. The other patterns show little evidence for deviations from expectancy on the global level. Table 4 gives details about the (only) $\langle\ \rangle$ pair, $LV1/\chi^2$. The upper two entries in the m-th box indicate the case giving rise to the m-th smallest p-value and that value itself, respectively. The third entry shows the value of C_m, which stays below 0.05 for $m = 1, \ldots, 5$. Agent #16 contributes the two smallest p-values.

Combining results globally might obscure systematic differences between the three intentions. Figures 2 to 4 show what happens if results are combined only across agents while keeping the intentions separate.

Comments. (i) For intention LO there is no sign of disagreement with $\mathcal{H}_0$. For intention HI there is only one $\langle\ \rangle$ pair, namely pattern *in* [97, 103], M test. The five smallest p-values stem from agents #11, 16, 2, 12, 7, in this order. It may appear strange that the two-sided M test yields more significant criteria than each of its one-sided variants M_+, M_-. This is due to the combination across cases and does not happen without combination; cp. Table 3. It indicates that in some of the m cases contributing to C_m the M_+ test is significant while for the others the M_- test is significant; or, put differently, the pattern occurs too frequently for some, and too rarely for the other cases. (ii) For intention BL there are six $\langle\ \rangle$ pairs and four $\langle\ \rangle$ patterns. Table 5 gives the details along the same scheme as in Table 4. For four of the six $\langle\ \rangle$ pairs, agent #16 contributes the

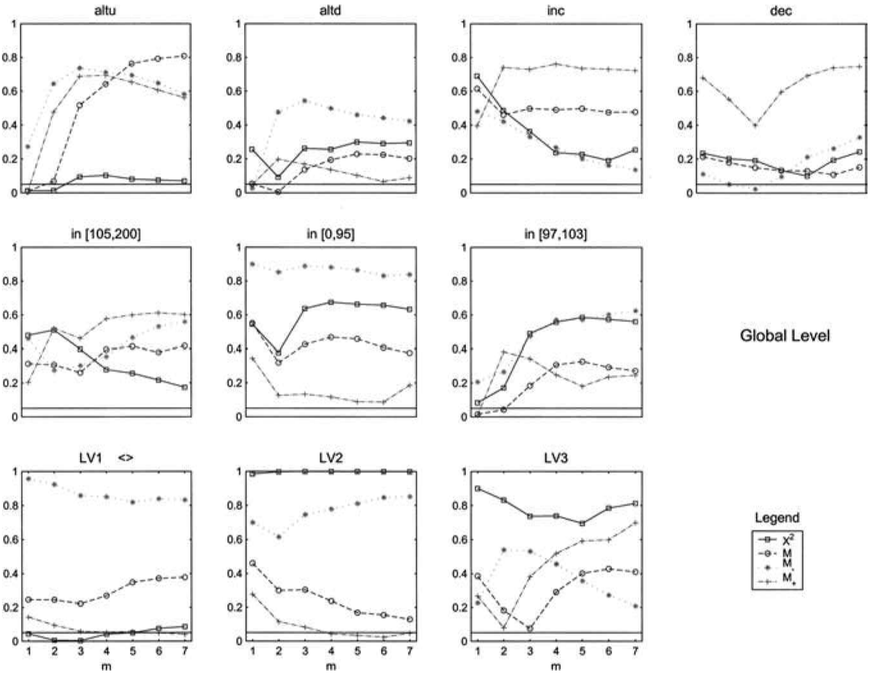

Fig. 1. Plots of criterion C against m, global level. Every subplot corresponds to one pattern and shows the results for the four tests. Tests coded according to legend.

smallest p-value. This fact should not be overstressed, however, because p-values for the various pattern/test pairs belonging to the same case are correlated. It would be more interesting if small p-values occurred with few agents only. The evidence is weak, though, e.g., *all* 16 agents appear somewhere in Table 5, and for each agent there is at least one pattern/test pair where the corresponding p-value is not among the seven smallest ones.

The results show that criteria differ considerably between intentions. The main effect, so to speak, is due to intention BL; intention HI contributes only one $\langle\ \rangle$ pattern (which is different from the BL ones), LO none. That may explain why on the global level only one $\langle\ \rangle$ pattern remains: no accumulation across

TABLE 4
Details about the $\langle\ \rangle$ pairs, global level. Upper entry: case giving rise to the mth-smallest p-value; middle entry: mth-smallest p-value; lower entry: value of C_{mr}

m	1	2	3	4	5	6	7
*LV*1	16/BL	16/LO	11/BL	7/HI	6/BL	11/HI	4/HI
χ^2	0.0009	0.0021	0.0084	0.0496	0.0504	0.0849	0.0959
	0.0432	0.0049	0.0035	0.0421	0.0493	0.0755	0.0861

TABLE 5
Details about the ⟨ ⟩ pairs, intention BL. Upper entry: agent giving rise to the *m*th-smallest *p*-value; middle and lower entries as in Table 4.

m	1	2	3	4	5	6	7
altd	2	14	15	4	13	8	3
M_+	0.0007	0.0214	0.0378	0.1003	0.1808	0.3446	0.3764
	0.0157	0.0304	0.0176	0.0280	0.0485	0.1174	0.1600
dec	16	2	13	6	12	4	3
χ^2	0.0053	0.0193	0.0478	0.1255	0.1348	0.1577	0.1720
	0.0849	0.0311	0.0231	0.0437	0.0371	0.0274	0.0176
dec	16	6	13	2	3	8	9
M	0.0050	0.0293	0.0430	0.0612	0.0656	0.2996	0.4063
	0.0765	0.0556	0.0273	0.0140	0.0055	0.0308	0.0767
*LV*1	16	11	6	8	5	13	10
χ^2	0.0009	0.0084	0.0504	0.1244	0.1524	0.2130	0.2589
	0.0146	0.0050	0.0151	0.0362	0.0394	0.0450	0.0464
*LV*1	16	11	10	8	6	7	1
M_+	0.0032	0.0122	0.0579	0.0739	0.0741	0.1062	0.1185
	0.0500	0.0128	0.0238	0.0171	0.0077	0.0043	0.0021
*LV*2	10	7	5	14	11	12	16
M_+	0.0106	0.0224	0.0446	0.1061	0.1228	0.2023	0.2164
	0.1562	0.0520	0.0276	0.0361	0.0294	0.0344	0.0296

intentions seems to take place, and "effects" rather get dissipated on the global level. Likewise, there is no clear evidence for a concentration on a few agents or cases, although case 16/BL stands out and agents #11, 7, 2, besides 16, appear to be well represented. A more careful discussion of this aspect could be interesting, but is not attempted here.

The predominant role of agent #16 prompts the question what happens if (s)he is removed from the analysis. It turns out that on the global level there is no pattern left fulfilling the ⟨ ⟩ pattern condition, and the same is true for intentions HI and LO. For intention BL there remain two ⟨ ⟩ pairs, namely *altd*/M_+ and *LV*2/M_+. It is interesting to look at Table 5 anew from this perspective. In summary, there is little evidence for possible deviations from $\mathcal{H}_0$ among agents #1–15, particularly if multiple testing is taken into account. On the other hand, the results for the case 16/BL resist a straightforward explanation by pure chance fluctuations.

4. Summary

The pattern count approach developed in this paper aims at complementing the omnibus type tests based on scaling index distributions by a number of statistically more tractable tests against more sharply defined alternatives. Ideally, it might indicate not only the existence, but the kind of a deviation from the null hypothesis in the time series under study. If the deviation is very small, as may be expected in the present context, then only a formal statistical test against a *predetermined, narrow* alternative can be expected to have the

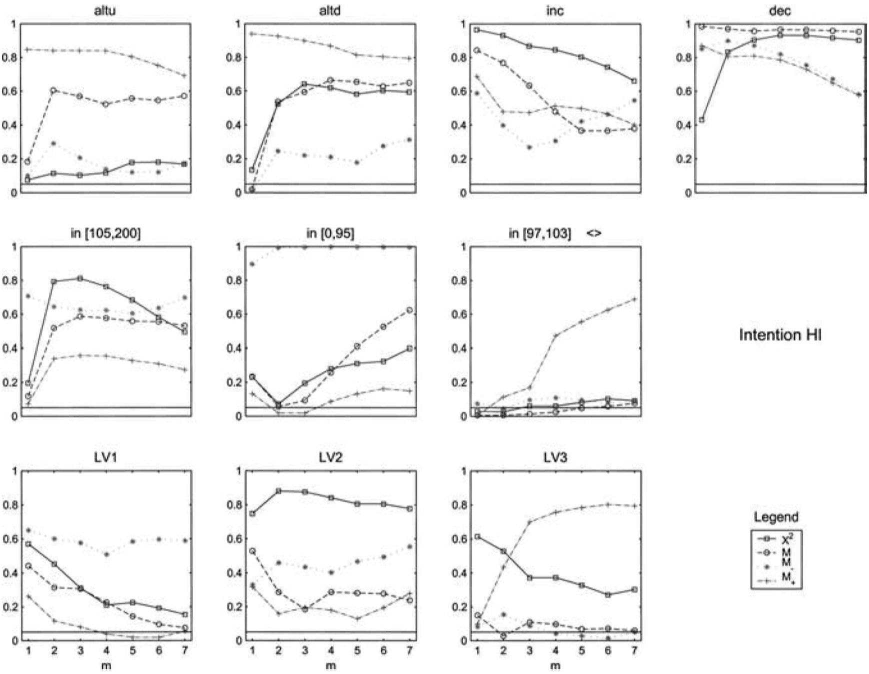

Fig. 2. Plot of criterion *C* against *m*, by intention. Otherwise as Figure 1.

necessary discriminatory power, at the cost of neglecting other alternatives. Our solution to this problem utilizes the fact that (pattern) count statistics are very sensitive to deviations from expected frequencies and simple enough to allow precise gauging of their fluctuations. On the other hand, the systematic variation of the patterns enlarges the range of detectable deviations.

In combining the individual results across agents and/or intentions, we assume that deviations from the standard random behavior of the RNG occur only sporadically and only with a few agents. Correspondingly we introduce a test criterion which combines improved detection power in such heterogeneous situations with the obligatory statistical validity. The criterion is defined as a sum of the *smallest* p-values associated with the individual cases, and is thus less affected by the "noise" contributed by many "no-effects cases" than the usual criteria which weigh all cases equally (and are, therefore, more efficient if deviations are homogeneous across the cases). In regard to multiple testing our strategy is to maintain the validity of statistical statements across various stages of data aggregation in order to provide a reliable basis for exploratory investigations at higher aggregation levels.

Our analysis of the data reveals various differences between the individual agents, intentions, and patterns. Patterns associated with persistence and trend-like features as *in* [,], *inc*, and *dec* are found conspicuous in a few cases only.

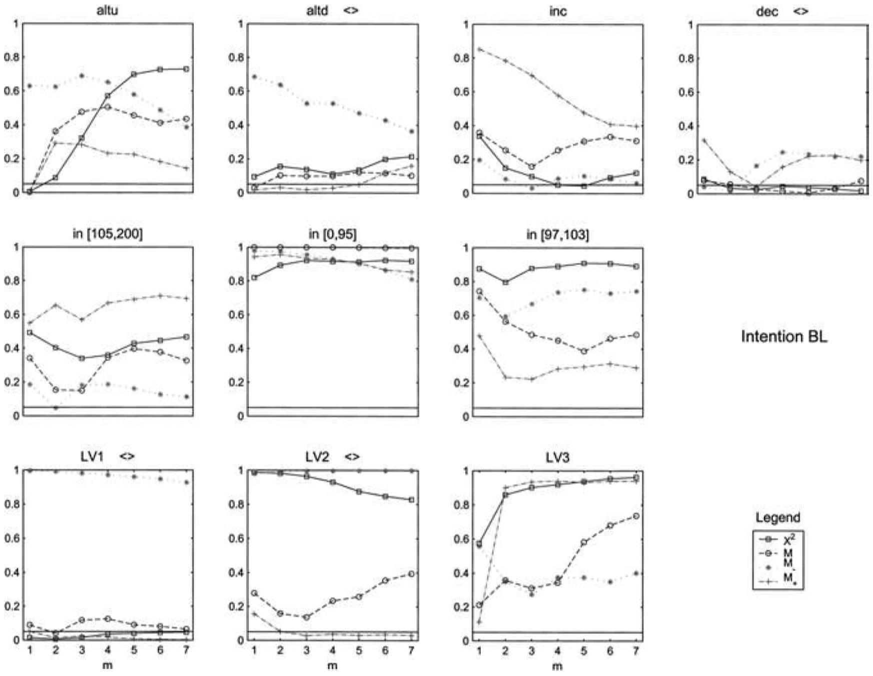

Fig. 3. Plot of criterion C against m, by intention. Otherwise as Figure 1.

Most of the (individually) significant results refer to the alternating and the local variation type patterns, which reflect oscillatory features of the time series. There is little evidence for deviations from $\mathcal{H}_0$ under the “active” experimental conditions HI and LO, whereas under the “passive” condition BL conspicuous results (conspicuous in the sense of “passing the $\langle\ \rangle$ condition”; cp. section 3.3) results are found for four among the 10 considered patterns. In view of the discrepancy between the conditions (intentions), one may expect that “effects” cancel each other if results are combined across intentions. Indeed, on the global level only one pattern remains conspicuous.

The very small individual p-values found particularly with the case 16/BL have to be interpreted with caution. Under the conditions of this study, individual p-values are fairly reliable down to order 10^{-3}. For even smaller orders, systematic and/or Monte Carlo fluctuation errors could happen to be of a larger order of magnitude than the p-value itself. Thus, there are also technical reasons to abstain from basing one’s analysis on *the* most extreme p-value, and to prefer instead a statistically more stable combination criterion like C.

The results for intention BL are largely due to agent #16; upon its removal from the data base the number of conspicuous patterns drops from four to two. Taken together with the fact that the agents contributing individually significant

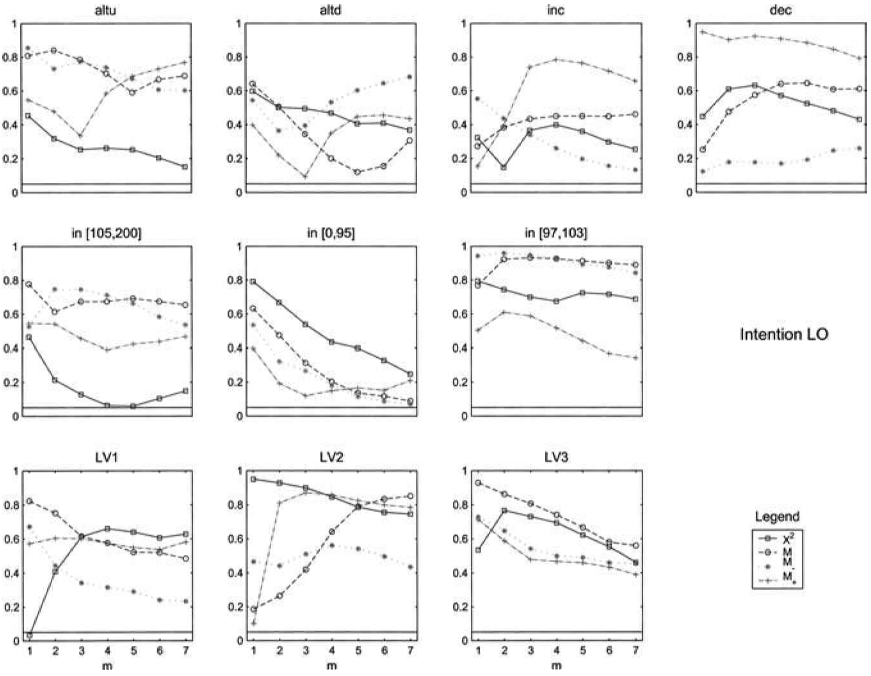

Fig. 4. Plot of criterion C against m, by intention. Otherwise as Figure 1.

results usually are different for different patterns, this indicates that the scattered, individually small p-values of agents #1–15 may be the result of chance fluctuations. On the other hand, the results for agent #16 are not easily explainable in this manner. This complies with and extends the findings of the scaling index analysis by Atmanspacher et al. (1999).

As emphasized previously, statistical confidence is not attached to these "higher level" conclusions, which are exploratory in character. As a matter of course the nature of possible causes which might explain the results of this investigation remains open.

Notes

[1] The comparison of different approximations to the exact null distribution of the p-values allows us to assess the quality of these approximations through internal consistency checks; see Appendix A.3. Such assessment is important because the validity of approximations in the tails of a distribution, which is the region of interest to us, often is questionable. Related errors may give rise to misleading results.

[2] On the first level one alternatively could combine results across intentions (and

report "by agent"). This route will not be followed, however, because our particular combination method will also convey information about "agent effect" even after aggregation across agents.

Acknowledgments

I am grateful to the MMI consortium directed by R. G. Jahn for making the data available for analysis. Thanks are also due to the reviewers and to H. Atmanspacher for careful reading of the manuscript and helpful remarks which improved the presentation of the paper.

References

Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer Verlag.

Atmanspacher, H., Bösch, H., Boller, E., Nelson, R. D., & Scheingraber, H. (1999). Deviations from physical randomness due to human agent intention? *Chaos, Solitons, and Fractals, 10,* 935–952.

Atmanspacher, H., Ehm, W., Scheingraber, H., & Wiedenmann, G. (2001). Statistical analysis of time series with scaling indices. *Discrete Dynamics in Nature and Society, 5,* 297–309.

Brown, B. M. (1971). Martingale central limit theorems. *The Annals of Mathematics & Statistics, 42,* 59–66.

Dobyns, Y. (1999). Harmonic analysis of cumulative traces. Preprint.

Jahn, R. G., Dunne, B. J., Nelson, R. D., Dobyns, Y. H., & Bradish, G. J. (1997). Correlations of binary sequences with pre-stated operator intention: A review of a 12-year program. *Journal of Scientific Exploration, 11,* 345–367.

Jahn, R. G., Mischo, J. Vaitl, D., et al. (2000). Mind/machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration, 14,* 499–555.

Karlin, S. & Taylor, H. M. (1975). *A first course in stochastic processes.* New York: Academic Press.

Pallikari, F. & Boller, E. (1999). A rescaled range analysis of random events. *Journal of Scientific Exploration, 13,* 25–40.

Reinert, G., Schbath, S., & Waterman, M. S. (2000). Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology 7,* 1–46.

Appendix: A Martingale Approach to Pattern Count Statistics

A.1 Basic Theory

Consider the observed time series $\{x_t\}$ obtained for a fixed agent/intention pair. Let $y_t = (x_{t-L+1}, \ldots, x_t)$, $t \geqslant L$, denote the vector of x-values observed at the L "time points" preceding and including "time" t. A *pattern* of length L, $L \geqslant 1$, will be any (measurable) subset B of R^L, the Euclidean space of dimension L. We say that *the pattern B occurs at time t* if $y_t \in B$, that is if the vector y_t falls into the set B. The frequency $S = S_\tau(B)$ of the pattern B in the segment x_t, $t = 1, \ldots, \tau$, can then be represented by the sum $S = \sum_{t=L}^{\tau} J_t$ of the indicator functions $J_t = (y_t \in B)$ taking the value 1 if $y_t \in B$, and 0 otherwise. [Below

$(y_t \in B)$ will denote both the event and its indicator function, the proper interpretation being clear from the context.]

For the sake of statistical analysis, the time series $\{x_t\}$ has to be conceived of as the realization of a stochastic process. In the present case it is assumed that $x_t, t = 1, \dots, \tau$, are (the realizations of) i.i.d. random variables on some underlying probability space $(\Omega, \mathcal{F}, P)$ with distribution $\mathcal{B}(200, 1/2)$. This assumption constitutes the null hypothesis $\mathcal{H}_0$; cp. the Introduction.

In theory, a test of $\mathcal{H}_0$ based on pattern frequency S can be constructed easily by referring the standardized counts ("z-score") $(S - \mu_0)/\sigma_0$ to a normal distribution. Here μ_0 and σ_0 denote the mean and the standard deviation of S under $\mathcal{H}_0$. In practice, depending on the nature of the stochastic process and the type of pattern, calculation of the moments μ_0 and σ_0^2 can become quite difficult, however, particularly if further complications arise due to discreteness of the x_ts, as in the cases studied herein.

Here we use the theory of martingales (a special class of stochastic processes; for a good introduction see Chapter 6 of Karlin and Taylor 1975) to circumvent the related problems. The basic idea is to center the variable S not at its expectation, by subtracting the (sum of the) unconditional probabilities $P(y_t \in B)$, but by subtracting suitable *conditional* probabilities which are easy to compute and make the difference a martingale. One thus can completely avoid calculation of the unconditional moments.

Technically this is accomplished by representing the set B as the Cartesian product of a subset B_α of R^{L-1} times a collection of subsets $B_\omega(\tilde{y}_t)$ of R^1 indexed by the first $L - 1$ coordinates $\tilde{y}_t = (x_{t-L+1}, \dots, x_{t-1})$ of y_t such that

$$y_t \in B \Longleftrightarrow \tilde{y}_t \in B_\alpha \quad \text{and} \quad x_t \in B_\omega(\tilde{y}_t). \tag{1}$$

B_α can be visualized geometrically as the projection of B onto the first $L - 1$ coordinates, and $B_\omega(\tilde{y}_t)$ as the intersection with B of the line parallel to the last coordinate passing through the "base point" $\tilde{y}_t$. Another interpretation closer in line with the present view is in terms of patterns: the pattern B occurs at time t if and only if it (i) "has been built up partially until time $t - 1$" and (ii) is actually completed at time t.

For every $t \geqslant 1$, let $\mathcal{F}_t \subset \mathcal{F}$ denote the σ-algebra generated by the random variables $x_1, \dots, x_t$, comprising all events which are observable until time t. Let $p_t = P[y_t \in B | \mathcal{F}_{t-1}]$ denote the conditional probability of the event $(y_t \in B)$ given the "past" $\mathcal{F}_{t-1}$. By (1) this can be written as

$$p_t = (\tilde{y}_t \in B_\alpha) P[x_t \in B_\omega(\tilde{y}_t) | \mathcal{F}_{t-1}]. \tag{2}$$

Finally, let

$$\Delta_t = J_t - p_t = (\tilde{y}_t \in B_\alpha)\{(x_t \in B_\omega(\tilde{y}_t)) - P[x_t \in B_\omega(\tilde{y}_t) | \mathcal{F}_{t-1}]\} \tag{3}$$

denote the conditionally centered pattern indicators, and $M_t = \sum_{s=L}^{t} \Delta_s$ their partial sums until time t. Then $(M_t, \mathcal{F}_t)$ is a *martingale,* i.e., the conditional expectations satisfy $E[M_t | \mathcal{F}_{t-1}] = M_{t-1}$ for every $t \geqslant L$ ($\mathcal{F}_0$ denotes the trivial

σ-algebra). Intuitively, processes with this property have no systematic tendency to increase or decrease over time. The martingale property thus expresses correct centering of S.

To normalize the variance of the conditionally centered counts we use the *predictable quadratic variation* defined by $V_\tau = \sum_{t=L}^{\tau} E[\Delta_t^2 | \mathcal{F}_{t-1}]$, which here reduces to a sum of conditional Bernoulli variances, $V_\tau = \sum_{t=L}^{\tau} p_t\,(1 - p_t)$. From the mathematical results described below it follows that for given pattern B the statistic

$$Z_\tau = \frac{M_\tau}{\sqrt{V_\tau}} = \frac{\sum_{t=L}^{\tau} [(y_t \in B) - p_t]}{[\sum_{t=L}^{\tau} p_t(1 - p_t)]^{1/2}} \tag{4}$$

can be used as a “z-score” for testing whether B occurs more (or less) frequently than expected under $\mathcal{H}_0$. In order to get some indication of how the difference between observed and expected pattern frequency develops over time one also may plot the *scores process* $Z_t = M_t/\sqrt{V_\tau}$ against time t.

So far everything referred to one single pattern. Now, suppose there are several patterns B^k, $1 \leqslant k \leqslant K$ (usually of a similar kind). Without loss of generality we may assume that the K patterns have identical lengths L. The quantities M, Z, etc. associated with pattern B^k will be characterized by the superscript k, and M_τ, Z_τ etc. now denote the column vector in R^K with components M_τ^k, Z_τ^k, etc. The analog of V_τ is the *predictable covariation matrix*, the $(k,\ l)$-th entry of which is

$$\begin{aligned} V_\tau^{k,l} &= \sum_{t=L}^{\tau} E[\Delta_t^k \Delta_t^l | \mathcal{F}_{t-1}] = \sum_{t=L}^{\tau} \{P[y_t \in B^k \cap B^l | \mathcal{F}_{t-1}] - p_t^k p_t^l\} \\ &= \sum_{t=L}^{\tau} \{(\tilde{y}_t \in B_\alpha^k \cap B_\alpha^l) P[x_t \in B_\omega^k(\tilde{y}_t) \cap B_\omega^l(\tilde{y}_t) | \mathcal{F}_{t-1}] - p_t^k p_t^l\}. \end{aligned} \tag{5}$$

Of course, all conditional probabilities and expectations in (5) have to be calculated with respect to the joint distribution of the random variables $\{x_t\}$ specified by $\mathcal{H}_0$.

For large τ the (unknown) distribution of the relevant statistics can be approximated by a normal distribution. Let *diag A* denote the diagonal matrix associated with a (square) matrix A, having all off-diagonal elements set to zero. For A positive definite and symmetric let us write $\lambda_{\min}(A)$ for the minimal eigenvalue of A, and $A^{-1/2}$ for the symmetric square root of its inverse A^{-1}. In the following we consider two different ways of standardizing the vector M_τ. Coordinate-wise standardization yields the vector $Z_\tau = (diag V_\tau)^{-1/2} M_\tau$, whose components are the test statistics (4) for the individual patterns. Joint standardization, $W_\tau = V_\tau^{-1/2} M_\tau$, yields a vector W_τ whose components are (asymptotically) uncorrelated. The symbol $\rightarrow_p$ denotes “convergence in probability”. Then the following holds.

Let I denote the $K \times K$ *unit matrix, and* $\Omega_\tau = EV_\tau$ *the (entry-wise) expectation of* V_τ*. Suppose that*

$$\lambda_{\min}(\Omega_\tau) \to \infty \quad \text{and} \quad \Omega_\tau^{-1/2} V_\tau \Omega_\tau^{-1/2} \to_p I \quad \text{as } \tau \to \infty. \tag{6}$$

Then the distribution of $W_\tau = V_\tau^{-1/2} M_\tau$ *converges (weakly) to* $\mathcal{N}_K(0, I)$*, the K-variate standard normal distribution. If furthermore* $(\text{diag}\Omega_\tau)^{-1/2}\Omega_\tau\,(\text{diag}\Omega_\tau)^{-1/2}$ *converges to a non-singular matrix* Γ*, then the distribution of* $Z_\tau = (diag V_\tau)^{-1/2} M_\tau$ *converges to the K-variate normal distribution* $\mathcal{N}_K(0, \Gamma)$ *with mean vector zero and covariance matrix* Γ.

A mathematical proof can be based on Brown's (1971) martingale central limit theorem and the Cramer-Wold device, or on Rebolledo's theorem as cited in Andersen et al. (1993). The conditions are fulfilled in the two important cases where the process $\{x_t\}$ is an i.i.d. sequence or an ergodic Markov chain with a countable state space. In order to avoid degenerate cases the events $(y_t \in B^k), 1 \leqslant k \leqslant K$ have to be *linearly* independent and their probability (taken under the stationary distribution in the Markov chain case) has to be different from 0 and 1.

The correlation between the empirical frequencies of the patterns (or the associated martingales) depends on their set-theoretic properties. The following two special cases mark the extremes. If two patterns B^1 and B^2 are mutually exclusive, $B^1 \cap B^2 = \emptyset$, then M_τ^1 and M_τ^2 are negatively correlated. They are positively correlated if one of the two patterns implies the other, $B^1 \subset B^2$, say. In fact, it is straightforward to verify that in the first case $V_\tau^{1,2} = -\sum_{t \leqslant \tau} p_t^1 p_t^2 \leqslant 0$, while in the second case $V_\tau^{1,2} = \sum_{t \leqslant \tau} p_t^1\,(1 - p_t^2) \geqslant 0$.

The application of martingale methods to counting processes is not new. For an overview emphasizing censored data analysis see Andersen et al. (1993). More closely related to the present work is the application to DNA analysis, reviewed in Reinert et al. (2000). There the patterns are words formed with a finite alphabet (A, C, G, T). Translated to the present situation this corresponds to patterns B consisting of one-element sets, which would be too sparse for the discretization level implied by a $\mathcal{B}(200, p)$ distribution. In principle, the above approach is applicable to every stochastic process. In applications the crucial point will be how difficult it is to calculate the conditional moments. Markov chains with known transition matrices are particularly well-suited in this respect. In our case no more is required than the calculation of binomial interval probabilities. These can be computed and stored in advance, making the appropriate centering and studentization of the count statistics numerically efficient.

A.2 Test Construction

The above result yields a variety of asymptotically valid tests by standard constructions of asymptotic statistical theory. The tests proposed here are of the

χ^2 and the maximum type. We refer to them as the χ^2 and the max test, respectively.

χ^2 test. This test rejects the null hypothesis if $\|W_\tau\|^2 = M'_\tau V_\tau^{-1} M_\tau > \chi^2_{\alpha,\,K}$, where $\chi^2_{\alpha,K}$ denotes the upper α quantile of a χ^2 distribution with K degrees of freedom (and x' the transpose of vector x).

max test. This test rejects $\mathcal{H}_0$ if the maximal absolute value $\|Z_\tau\|_\infty = max_k |Z_\tau^k|$ of the individual z-scores exceeds the critical value ζ_α. The latter is determined by the equation $P[\|\tilde{Z}\|_\infty > \zeta_\alpha] = \alpha$, where $\tilde{Z}$ is a random vector with distribution $\mathcal{N}_K(0, \Gamma)$.

It has to be emphasized that the *exact* distributions of the test statistics under $\mathcal{H}_0$, especially the exact critical values, are *unknown.* Therefore use of approximations (as the standard χ^2 approximation in case of the χ^2 test) is unavoidable, though possibly problematic. We shall come back to this issue in Appendix A.3 below. For the *max* test, not even an approximation is readily available: First, the matrix Γ is unkown, being defined in terms of the unconditional expectation Ω_τ of the predictable covariation matrix V_τ (and only in the limit $\tau \to \infty$). Secondly, even for known Γ the distribution of $\|\tilde{Z}\|_\infty$ is difficult to compute.

The most radical solution to these problems is to do away with asymptotics and to entirely rely on Monte Carlo simulation, as follows. One generates ν copies of the underlying time series (according to the distribution specified by $\mathcal{H}_0$) and computes the quantity $\|Z_\tau\|_\infty$ anew for each such copy. Unfortunately, however, for long time series (large τ) and the large ν required to get stable estimates this procedure becomes very time consuming. Therefore we propose a second, more efficient hybrid simulation method relying partly on asymptotic normality. For ease of distinction we refer to the first and the second simulation method as *fullsim* and *hybsim,* respectively.

Method *hybsim* consists of two steps. In the first step one obtains a data-based estimate of the correlation matrix Γ. For every agent/intention pair one computes the predictable covariation matrix V_τ from the corresponding time series, forms the average $\bar{V}_\tau$ of these matrices, and takes $\hat{\Gamma}_\tau = (\mathrm{diag}\bar{V}_\tau)^{-1/2}\bar{V}_\tau(\mathrm{diag}\bar{V}_\tau)^{-1/2}$ as the estimate of Γ. Under the conditions of the theorem, this estimate is *consistent,* i.e., it approximates Γ with high probability for large τ under $\mathcal{H}_0$, which is sufficient from the viewpoint of asymptotics. It should be noted that the consistency does not hinge upon *averaging* the individual V_τ matrices; theoretically, Γ is consistently estimable from every single time series as $(\mathrm{diag}V_\tau)^{-1/2}V_\tau(\mathrm{diag}V_\tau)^{-1/2}$. Nevertheless, the stabilization of the estimate of Γ due to averaging the V_τs should be advantageous.

In the second step one generates a large number μ (we took $\mu = 10^5$) of $\mathcal{N}_K(0, \hat{\Gamma}_\tau)$-distributed random vectors $\tilde{Z}_n$. Then one estimates the critical value ζ_α for a given level α as the empirical upper α quantile of the surrogate data $\|\tilde{Z}_n\|_\infty$, $1 \leqslant n \leqslant \mu$. Note that simulating vectors $\tilde{Z}_n$ is much more simple and fast than simulating complete time series and calculating the relevant statistics each time.

One-sided versions of the maximum test are of interest, too. Here one rejects $\mathcal{H}_0$ for large values of the statistics $\max_k Z^k_\tau$ or $-\min_k Z^k_\tau$, respectively. For these, Monte Carlo critical values are obtained in the same manner as above.

A.3 Assessment of Approximation Errors

The purpose of this section is to discuss sources and consequences of possible errors in the computation of p-values. Basically, there are two sources related to Monte Carlo fluctuation errors and to systematic biases due to asymptotic approximations, respectively.

Let us first consider fluctuation errors. The *fullsim* method is based on ν Monte Carlo copies of the time series $\{x_t\}$. Applied with any (continuously distributed) statistic T it yields an approximate p-value having distribution $\mathcal{U}_\nu$, the discrete uniform distribution on the grid of all numbers j/ν, $0 \leqslant j \leqslant \nu$. Since $\mathcal{U}_\nu$ approximates $\mathcal{U}$ arbitrarily well for large ν, this method is as good as one can possibly hope for. Nevertheless, it has its limitations. First, the distribution $\mathcal{U}_\nu$ obtains exactly only if the *fullsim* method produces an ideal i.i.d. sample $T^*_1, \dots,$ T^*_ν from the null-distribution F_0 of T, that is, if one disregards the possible problems due to, e.g., discreteness of F_0, biases of the pseudo random number generator, or numerical inaccuracies. Secondly, and more importantly, $\mathcal{U}_\nu$ is the *unconditional* distribution of $\hat{p}$, where both the Monte Carlo sample $T^*_1, \dots, T^*_\nu$ and the statistic T are considered as (i.i.d.) random variables. This fact by itself conveys nothing about the fluctuations appearing in repetitions of the simulation procedure, hence about its reliability. For a tentative assessment consider the observed value of T as fixed. Conditionally on T, $\nu\hat{p}$ has the binomial distribution $\mathcal{B}(\nu, p)$, where $p = 1 - F_0(T)$ is the "exact" p-value. Therefore, the magnitude of the fluctuations of $\hat{p}$ about p is of the order $\sqrt{p(1-p)/\nu}$, and, for $p < 1/2$ and relatively with respect to p, of the order $(\nu p)^{-1/2}$. This relative magnitude is small if the expected number νp of Monte Carlo variables T^*_j exceeding T is large; otherwise, that is for p-values $p \ll 1/\nu$, the fluctuations become substantial. Practically speaking this means that if in 50000 Monte Carlo samples, say, there are 5 exceedances then the *fullsim* p-value of 10^{-4} is subject to considerable chance fluctuations. With 50 exceedances the p-value 10^{-3} is already rather more reliable.

Normal or χ^2 approximations $\hat{F}_0$ suggested by large sample theory induce a systematic (deterministic) bias, which is further overlain by random errors if large sample theory is combined with simulation as in the *hybsim* method. A comparison of the *hybsim* and *fullsim* p-values for all cases and all pattern/test pairs showed that relative errors are often below 0.1, and above 0.3 only in few cases and only for p-values close to 0 or 1. The largest relative error, 4.0, occurred for the configuration "case 11/HI, pattern *in* [97, 103], M_+ test" and the p-values 0.00020 (*fullsim*) and 0.00004 (*hybsim*). (Note that a p-value of 0.00004 estimated from 50000 Monte Carlo simulations corresponds to just two nonrejections of $\mathcal{H}_0$!) The second largest was 0.72 (same configuration, apart from M instead of M_+ test).

Since the combination criterion C depends on selection of the m smallest p-values, it is possible that one gets different sets of m associated cases if the p-values are determined by different approximation methods. To check this we ordered the cases by increasing p-value and systematically searched for instances where the induced ordering depends on the p-value approximation method. An instance where this did happen was for pattern *in* [0, 95], intention HI (aggregation level "by intention"): the 6th and 7th smallest of the M_+ test p-values stem from agents #8 and #9, respectively, if computed by *hybsim,* whereas if computed by *fullsim* they stem from agents #9 and #8. The associated p-values ranged from 0.29058 to 0.29411. For the M test, too, the ordering of agents #8 and #9 was permuted for the *fullsim* p-values compared with the *hybsim* ones. (Here the p-values were the 7th and 8th smallest and ranged from 0.55983 to 0.56316.) In all other instances the ordering of cases did not depend on the approximation method.