Journal of Scientific Exploration, Vol. 31, No. 2, pp. 231–254, 2017 0892-3310/17

RESEARCH ARTICLE

Multiple-Analysis Correlation Study between Human Psychological Variables and Binary Random Events

Hartmut Grote

Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), and Leibniz-Universität, Hannover, Germany
hrg@mpq.mpg.de

Submitted July 30, 2016; Accepted February 22, 2017; Published June 30, 2017

Abstract—Mind–matter interaction experiments have been progressing from targeting simple bias of random number generators to correlation studies between psychological and physical variables, carried out over multiple combinations of these. This paper reports on a new correlation study between human intention and the output of a binary random number generator. The study comprises a total of 720,000 bits from 20 equal sessions, each with a different human participant. Each participant spent one hour attempting to 'influence' the outcome of the random number generator according to a pre-selected intention. During this time the participant was provided feedback on his/her performance by an analog mechanical display, with the needle of a galvanometric instrument moving to the left or right of its initial position, according to the instantaneous output of the random number generator. Psychological variables were obtained from the participants by a hardware dial ahead of each individual run and by a questionnaire before the participant's first experimental session. Three types of data analysis were defined and tested before looking at the data, resembling a blind analysis technique. The first analysis looks at the distribution of hit rates from the 20 participants. A former study of this kind had found a significant result for this type of analysis (Grote 2015). The second analysis tests for correlations between psychological variables obtained before each run and the hit rate of the corresponding subsequent run. The third analysis is a conceptual replication of von Lucadou's correlation matrix method. It consists of multiple correlation tests between psychological and physical variables, which also can be interpreted as a multiple-analysis technique. The results of the study are p-values of $p = 0.438$, $p = 0.703$, and $p = 0.0949$ for the three analysis' results to have occurred by chance under a null hypothesis. The combined p-value for these results is $p = 0.315$. While none of the pre-defined analysis results is significant, a post hoc variant of Analysis 3 that includes the control data is significant with $p = 0.012$ to have occurred by chance, under a null hypothesis.

Introduction

The debate on the existence or non-existence of mind–matter interaction (MMI) is a topic at the fringes of mainstream science, with sometimes strong opinions held by individual researchers defending either view. While for some researchers in the field of anomalous psychology, the existence of mind–matter interaction seems beyond doubt (see, e.g., Radin & Nelson 1989, 2003, Jahn & Dunne 1986), this is not the case at all for the majority of the scientific audience (Odling-Smee 2007, Bösch, Steinkamp, & Boller 2006). Experimental evidence is often a matter of the interpretation of the studies, which makes it difficult for new researchers to form an opinion on the research performed to date, as is visibly exemplified in the dispute on the interpretation and validity of meta-analysis of existing mind–matter experiments (Bösch, Steinkamp, & Boller 2006, Radin et al. 2006, Pallikari 2015). See also the references in Bösch, Steinkamp, and Boller (2006) for an overview of existing research.

Also, the more cautious label of mind–matter correlation (i.e. correlation between human intention and the output of a physical system), which may not postulate direct causality, seems largely neglected by most scientists, even though attempts at explanations of a putative correlation effect, like for example the interpretation as entanglement correlations in a Generalized Quantum Theory (Atmanspacher, Römer, & Walach 2002, Filk & Römer 2011) do exist (von Lucadou, Römer, & Walach 2007, Walach, von Lucadou, & Römer 2014).

For these reasons, it seems of some value to the field if new mind–matter experiments are performed from time to time, in particular if new researchers are involved in conducting such experiments and possibly new aspects are introduced in the experimental approach. The latter should also serve to prevent strict replications of earlier MMI-like experiments, which may suffer from a possible decline of a putative effect, found by a number of replication studies in this field, and discussed for example in Kennedy (2003), von Lucadou, Römer, and Walach (2007), Walach, von Lucadou, & Römer (2014), and references therein.

The study described in this paper is the second study by this author. The first study is described in Grote (2015), and the experimental setting of that study has been modified in the following ways:

1. The rate of random bits produced has been reduced from 1,000 bits/s to 10 bits/s.
2. The random bit generation process has been modified from a Schmidt process to a 1-step Markov process.[1]
3. The feedback has been extended to include a color-lighting

scheme in the background of the galvanometer needle.
4. Feedback has also been extended by the sound of a gong, which is played during a run, if the participant is successful.
5. The sequence of left/right intentions is recorded in the new experiment.
6. The duration of a single run has been reduced from 60 s to 30 s.
7. Before each participant starts the first run, psychological variables categorized into 6 items have been obtained by questionnaire.
8. Before the start of each run, psychological variables are obtained from the participant.
9. The number of participants is 20.

Items 1 and 2 have been introduced based on a suggestion by W. von Lucadou. Items 3 and 4 have been introduced to potentially increase the focus of the participants, and items 5 to 8 allow for different types of analysis, mostly searching for correlations between psychological and physical variables.

While the outcome of the first study (Grote 2015) was not significant overall, one out of four individual analyses was found significant. That analysis is also carried out in this study (Analysis 1), testing the distribution of the basic results (*z*-scores) of the 20 participants. Analysis 2 in this study tests for correlations between three psychological variables obtained *before each run*, and the basic outcome (number of hits above chance expectation) of the corresponding runs. Analysis 3 is a conceptual replication of the correlation matrix technique that has been used by von Lucadou and others (von Lucadou 2006, von Lucadou, Römer, & Walach 2007), though with fewer variables and fewer participants. This technique uses multiple correlation tests between physical variables (properties of the data) and psychological variables (properties of the participants). No predictions are made about which of the correlations would be significant, but rather the combined significance of all correlations is assessed. This is further detailed in the section ***Pre-planned Data Analysis***.

The analysis of the data was defined and tested before any of the data were actually analyzed, which is also referred to as a *blind analysis* method (Klein & Roodman 2005). Blind analysis is a strict form of a pre-specified analysis in which the analysis code is fully implemented and tested before the data are looked at. Blind analysis is particularly useful in looking for small effects in noise and, in the opinion of the current author, is well-suited to address criticisms of data analysis (Wagenmakers et al. 2015) and of questionable research practices (Bierman, Spottiswoode, & Bijl 2016) in this domain of research.

It was decided to attempt to publish the result of this study regardless of the outcome of the analysis, in order to not contribute to publication bias.

In the section ***Experimental Design***, the experimental setup is described, followed by the section ***Pre-planned Data Analysis*** on the pre-defined data analysis plan. The results of the analysis are presented in the **Results** section. Finally, the **Discussion** section contains a brief discussion of the analysis and results.

Methods

Experimental Design

The experiment described in this paper was designed and conducted by the author. Participants were 20 people (including the author) in different relationships with the author (i.e. friends, friends of friends, work colleagues, etc.) who were interested in the topic, and willing to spend one hour each on actual experimentation time. The participants' age spanned from 21 to 76 years old with a mean age of 46 and a standard deviation of 13 years. The participants included both genders, 11 female and 9 male.

Each participant had agreed to carry out 120 "runs" of the experiment, with each run lasting 30 seconds. A single run would always begin by the participant selecting whether he/she would try to influence the motion of the needle of a galvanometer display to the left side or to the right side during that run. This choice had to be executed by the participant by pushing a switch either to the left or to the right, respectively. The chosen direction would then be displayed to the participant throughout the following (30-s long) run, in order to remind the participant of the chosen direction.

Next, the participant had to turn a dial in order to choose on a scale from 0 to 10 his/her actual mood (0 meaning 'very bad mood', 10 'very good mood'). This dial consisted of a rotary knob that could be rotated by about 270 degrees, in order to choose a number between 0 and 10, which would be displayed to the participant while the knob was rotated. Then the participant would press the 'start' button to begin the 30-s long run. While the run was active, a colored light was lit in the background of the display needle, to signal to the participant that the run was in progress. Figure 1 shows a photograph of the galvanometer display with the background lit during a run.

During each 30-s long run, random binary events would be generated at a rate of 10 per second. A Markov chain with a memory length of one was used to generate the random numbers, as described below. The draw (from the Markov chain) of a logical '0' would result in a step of the display needle to the left side of its current position, while a logical '1' would result

Figure 1. The galvanometer display with lit background during a run.

in a step of the needle to the right side of its initial position. In this way, 300 binary random draws were accumulated during each 30-s run, resulting in a corresponding random walk of the needle. The maximal range of the needle was 11 steps in either direction, with one standard deviation equal to 5 steps ($N = 300$ for Equation 2 below). The color of the light in the background of the display needle was made to change in correspondence with the position of the needle. Additional feedback was given to the participant by playback of a gong sound when the participant exceeded a threshold of 6 steps in the intended direction over the expectation value (zero steps), during the ongoing run.

The participants operated the device (almost exclusively) at their homes and at times convenient to them, according to their own choice. They were instructed to if possible be alone in the room when operating the device, and to finish the assigned 120 runs within one to two weeks if possible.

An individual run of 30 s could not be interrupted by any means, by an internal mechanism that inhibited switching the device off while a run was proceeding. An internal battery in the device assured that the device would run independent from the main power and thus also independent from any possible interruption of the main power during a run. The participants were free to distribute the time to perform the runs at their choice and could choose for any run between left or right intention, but had to respect the constraint that over the 120 runs both left and right intention had to be picked the same number of times, 60 each, respectively. For example, it would have been possible to do all 60 left-intention runs first, followed by the 60 right-intention runs, but the device would not allow for either intention to be chosen more than 60 times, to assure the balancing of

intentions. Therefore, each participant conducted 60 runs with left intention and 60 runs with right intention, accumulating one hour of data in total. Each participant committed to collect this one hour of experimental data, and each participant fulfilled this goal. The total timespan used by the participants to complete the 120 runs varied from less than 1 day to about 4 weeks. The experimental data-taking started in the spring of 2014 and concluded in the summer of 2015, when the number of 20 participants had been reached. Up to four participants could share the device (e.g., members of a family) by freely distributing experimentation time among themselves. Each participant simply had to choose his/her name on the display ahead of a run, in order to allow the data to be associated with the correct participant.

The data of the experiments were stored in two different formats in the device as a safeguard against data errors. No such errors occurred. The data were transmitted to a personal computer after 1 to 4 participants had completed their runs. This data transmission used check-sums to safeguard against transmission errors, and no such errors occurred.[2] The device was then prepared for the next participant(s) by resetting the data memory of the device and programming the names of one or more new participants.

In addition to participant data, a set of control data was taken, which was not explicitly subject to any interaction with the intention of any participant.

Between participants (i.e. when the device was in the hands of the conductor of the study for transferring data and preparing the device for new participants), a number of complete datasets for 'dummy participants' were automatically generated. For this purpose, dummy persons with names '01' to '20' were generated by the conductor, and when the device would recognize a dummy participant name (by the fact that such a name would start with a *number* rather than with a *letter*), it would automatically start an individual run after a random time interval of order 1 minute length. The 'intention' for each such run was chosen randomly by the internal hardware random number generator (RNG) (see the section ***The Binary Random Number Generator***) but satisfying the required equal total number of left and right intentions as for the real runs. This way a complete set of 20 dummy participants was created, spread throughout the time of acquisition of the participants' data, which is taken as a complete control dataset for the study.

As a particular feature of this study, the participants carried the experimental device to their homes, where they could 'work' on the experiment, at the time and in the environment of their choice. While this may appear to be giving up control over the conductance of the experiment compared with a laboratory setting, it has the advantage that the participants might feel more at ease in environments of their choice, and thus might get

more involved in their effort to 'influence' the needle. Ultimately, even in the laboratory, the conductor of the experiment has no control of whether the participant would assert 'influence' on the device according to the pre-stated intention or not. Although no fraud on the participants side was to be expected whatsoever, principal measures to detect physical manipulation or malfunctioning of the binary random number generator were taken, as detailed below.

The author preferred to choose a real physical system (the needle of a galvanometer display) over a computer screen, which is often used in other experiments of this kind. Computer screens are so common in our modern life, that a mechanical display carries the element of 'being different'.

A description of the random number generator is given in the Appendix section ***The Binary Random Number Generator***.

Pre-Planned Data Analysis

To avoid bias, the data analysis procedure was defined and tested before any of the data were actually looked at. Three different investigations (named Analysis 1, Analysis 2, Analysis 3) were carried out, as described in the following subsections. The principal outcome of each of the three analyses is a number describing the probability that the obtained result would have occurred by chance under the null hypothesis, i.e. assuming no correlation between the data and experimenters' intention.[3] The chance probability for the combined results of the three investigations is also given.

Each of the 3 analyses uses simulated (Monte Carlo) data, in order to estimate likelihoods of test results from the participants (and control) data. Using simulated data is a standard technique when the background cannot be easily modeled analytically and in low–signal-to-noise experiments. The null-hypothesis distributions against which the measured scores are evaluated are generated using software random number generators, simulating trials like the ones that the participants in the experiment undertake. However, there is actually no participant providing an intention and so we take the results from these fake-trials as expressions of the statistical scores under the null hypothesis.[4] The simulated (Monte-Carlo) data consist of 10,000 complete sets of data, each resembling data of a full study comprising 20 'participants'.

Another feature of the analysis is that in particular Analysis 2 and Analysis 3 have several degrees of freedom, which is equivalent to applying several tests to a set of data. However, no predictions are made about the outcome of individual tests, but the results of a number of tests are combined into one 'figure-of-merit' (FOM), which can also be called a 'test statistic'. This FOM can, for example, be the product of the estimated likelihoods

of individual test results. This principle was inspired by the correlation matrix technique used by von Lucadou and others, as mentioned in the **Introduction**. In the form used here in Analysis 2, it mainly consists of a method to perform multiple analysis. Analysis 3 is a conceptual replication of the correlation matrix technique as detailed below.

The control dataset, as defined in the above section ***Experimental Design***, will be subject to the same Analyses (1, 2, and 3) as the main dataset. However, the control data play no role in the pre-defined analysis, and can be viewed as a consistency check or can be used in post hoc analysis. Since for the control data there exists no separate set of psychological variables, the psychological variables of the 20 participants are used to be correlated with the control data (this applies to Analysis 2 and Analysis 3, where psychological data are used for correlation with physical data).

All three analyses have been tested with fake datasets, which have been generated by an independent (independent from the algorithm used to generate the simulated/Monte Carlo data) algorithm. No deviation from the expected uniform distribution was found in the 100 datasets used for testing.[5]

Analysis 2 and Analysis 3 have also been tested with dedicated fake datasets that included intentional biases tailored to the specific analysis. This way the proper functioning of the analysis was confirmed, i.e. the ability to detect what the analysis is supposed to detect.

Finally, we point out that the description of the experiment, the definition of the pre-planned data analysis, as well as the analysis code and the complete experimental data, have been uploaded to the website *opensciencefframework* (https://osf.io/) prior to the actual analysis of the data. Also prior to the actual analysis, the data on said website were marked as a read-only representation of the project (i.e. it cannot be modified anymore), and can be made accessible upon request to the author. In particular, this procedure is a *blind analysis* procedure. The pre-defined analysis of the experimental data is performed only after the analysis code has been frozen. In principle, it can then be performed by a single button press. This process is called the *unblinding* or *opening of the box* in other fields. Blind analysis has been successfully applied in nuclear physics and particle physics (Klein & Roodman 2005) and is the standard method to analyze data in these fields today.

Analysis 1. We define a hit to be a high bit when the participant's intention was to move the needle to the right, and to be a low bit when the participant's intention was to move the needle to the left. The total number of hits n_{hits} is the sum of hits scored under *right* intention plus the hits acquired under *left* intention. The z-value over a total number of trials N is then defined as

$$z = \frac{n_{hits} - N/2}{SD} \tag{1}$$

The standard deviation SD is estimated as

$$SD = \sqrt{N/12} \tag{2}$$

Note that the factor 12 under the square root comes from the fact that we obey the statistic of a 1-step Markov chain (von Lucadou 2006), where each random bit depends on the last random bit, as a result of the bit-generating procedure described in the Appendix section ***The Binary Random Number Generator***. The z-score is a useful quantity because it provides an immediate sense of the deviation of the results from expectation.[6]

For Analysis 1, the data as detailed above (z-scores for the number of obtained hits) are calculated for each of the 20 participants separately, such that 20 z-scores are generated. These 20 z-scores are then sorted and (frequentist) p-values are generated for the highest ranking, second-highest ranking, third-highest ranking, and so forth down to the lowest ranking, by comparison with the distribution of the same ranking values determined from a simulated (null hypothesis) dataset. These p-values are two-sided, with $p = 1$ if a data point is exactly in the middle of the compared distribution. The resulting 20 p-values are combined (by summing over the inverse squares of p-values) and result in the figure of merit (FOM) for this test. The chance probability for the value of this FOM is measured against the distribution for the same FOM derived from the Monte Carlo dataset. A one-sided probability will mean that the FOM of the test data (or a lower one) has occurred by chance. This is the result of Analysis 1.

Notes on Analysis 1. This analysis is sensitive to the *distribution* of results among the participants. It is also sensitive to deviations from randomness in directions opposite to a participants' intention. No prediction is made on how in particular the individual results would deviate from the expected distribution. However, a one-sided probability is chosen as the main result, under the hypothesis that deviations would more likely show up in the direction of deviations of individual results from their reference class. A probability of this analysis that is close to unity would indicate that the participants' data are closer to the expected distribution than expected by chance.

The total hit rate over all participants, which is the classical type of analysis for this kind of experiment, is not foreseen as a test, but can be considered as post hoc analysis, while explicitly not counting in the final statistical evidence of the study at hand.

Analysis 2. This analysis comprises three correlation tests between three different psychological variables obtained before each run, and the hit rates of individual run results.

The three psychological variables used are:

- The variable *mood*, obtained before each individual run on a scale of 0 to 10.
- The variable *time*, also obtained before each individual run, which is the time the participant needed from starting to choose the mood parameter (by turning the mood dial) to the actual start of the run (by pressing the start button).
- The variable *sequence*, which is a measure of how many runs in the past the direction of the intention (left or right) was chosen to be the same as for the actual run.

For the calculation of the correlations, Spearman's rho is used. The correlations are split between right and left intention, such that there are two correlations calculated for each psychological variable (and for each participant). Each correlation uses the 60 hit rates (as defined in the subsection **Analysis 1** of the ***Pre-Planned Data Analysis*** section) for each run of either left or right intention. The p-values of the two resulting correlation factors pertaining to one psychological variable are multiplied and yield the test result for one correlation test. This procedure is performed for all 20 participants, and the 20 test results are multiplied to yield one combined result for each psychological variable.

Each of the three combined results is then compared to the equivalent test results of a large number of simulated data (again by a ranking). By this comparison, a two-sided (frequentist) probability is estimated for each test, that the acquired result (or a lower/higher one) would have occurred by chance. In a second step, all of these probabilities (one for each statistical test) are combined (by summing over the inverse squares of p-values) to yield a single figure of merit (FOM) of the acquired data. Finally, this FOM is compared to the distribution of the same FOMs of the simulated data, and a one-sided (frequentist) likelihood results, that the actual FOM (or a lower one) of the data under test would have occurred by chance. This likelihood is the result of Analysis 2.

Notes on Analysis 2. While basically a test of 3 correlations, this analysis can also be interpreted as a correlation matrix technique as described for example in von Lucadou (2006). A correlation matrix (as used in these references) shows the number (and strength) of correlations between several physical and psychological variables of the experiment as a whole. In terms of Analysis 2 defined here, there are 3 psychological variables, and one

physical variable, such that this 'matrix' has only three entries. However, one could also argue that three correlations are actually calculated for each participant, which are then combined for all 20 participants. In this sense we have 60 correlations.

Analysis 3. This analysis is a conceptual replication of the correlation matrix technique used by von Lucadou and others.

Psychological variables of each participant have been obtained by questionnaires before the start of the first run of that participant. The questionnaires are summarized into the following categories, to form 6 psychological variables:

- TAS: Tellegen absorption scale with 34 items
- SG: Sheep–Goat scale with 9 items
- SENS: reduced sensitivity person scale with 9 items
- TRANS: reduced transcendental scale with 6 items
- EX: Extraversion scale with 12 items
- MED: Experience with a meditation technique

Five physical variables are formed for each participant, resulting from the 120 runs that each participant conducted:

- HIT: Total hit rate
- ACR: Autocorrelation of the time series data, shifted by 1 and 2 s
- RUN: Runtest of time series data, testing the hypothesis that the data are randomly distributed in time
- EXC: Number of excursions in intended direction
- GNG: Number of audio feedbacks (gongs) obtained

For each psychological variable, the correlation with each physical variable is calculated using Spearman's rho. The resulting 30 values are then compared to the equivalent test results of a large number of simulated data (again by ranking). By this comparison, a two-sided (frequentist) probability is estimated for each test, that the acquired result (or a lower/higher one) would have occurred by chance. In a second step, all of these probabilities (one for each correlation) are combined (by summing over the inverse squares of p-values) to yield a single figure of merit (FOM) of Analysis 3. Finally, this FOM is compared with the distribution of the same FOMs of the simulated data, and a one-sided (frequentist) likelihood results, that the actual FOM (or a lower one) of the data under test would have occurred by chance. This likelihood is the result of Analysis 3.

Notes on Analysis 3. This is the first independent conceptual replication of the *correlation matrix method* (CMM) using multiple participants. A brief explanation of the CMM method can be found in the Appendix section ***Notes on the Correlation Matrix Method***.

A strange anecdotal occurrence: Trickster at play? The author (*I*, for this section) would like to share an anecdotal occurrence here, which happened during the testing of the data-analysis procedures. There are 3 types of analysis defined, as described above. As far as my best memory goes, on all three occasions of *first* testing each analysis (but certainly for 2 of them), the very first statistical outcome for a single test dataset was rather on the edge of the distribution of possible outcomes (of order 1% or lower), which initially raised my concern with the validity of the analysis. However, after applying more than 100 test datasets, the statistics of the outcome resolved to the expected normal distribution, for all 3 analyses, as stated above. For all of these tests, new test data were generated and the system timer value was set as seed number to the pseudo-random algorithm before generating each test dataset.

One other similar instance happened with an auxiliary analysis for Analysis 3 which was testing the counting method of matrix elements above a threshold, rather than using the pre-planned method of combining all matrix elements. When first testing the counting of correlations above a threshold, again with a fresh set of simulated data, on the very first instance this number was found to be 8. According to the test with many hundred simulated datasets afterward, the likelihood for obtaining 8 significant results is about 0.1%. Just to be clear, the generation of the matrix correlation factors was not changed on this occasion, just their evaluation via the threshold method was tested as an auxiliary investigation of the analysis procedure.

Taking at least 2 instances with 1% chance and one with 0.1%, this gives a combined chance of about 10^{-5} using Fisher's method for combining p-values uniform on the interval [0,1] (Fisher 1970). Of course, this was not predicted, and is a spontaneous observation, which, however, I found quite curious and which reminded me of G. Hansen's book *The Trickster and the Paranormal* (Hansen 2001) as well as J. Kennedy's paper "The capricious, actively evasive, unsustainable nature of psi" (Kennedy 2014).

It is obvious that the testing of the pre-defined analysis with a single set of test data can be viewed as a PK-like experiment on its own. The Trickster quality of this occurrence is interesting to contemplate.

Results

Analysis 1

Figure 2 shows the result of Analysis 1. The probability of the participants' results to have occurred by chance (null hypothesis) is $p = 0.438$, which is not significant. This probability is obtained by the fraction of more extreme results (more negative FOM) divided by the number of all results of the

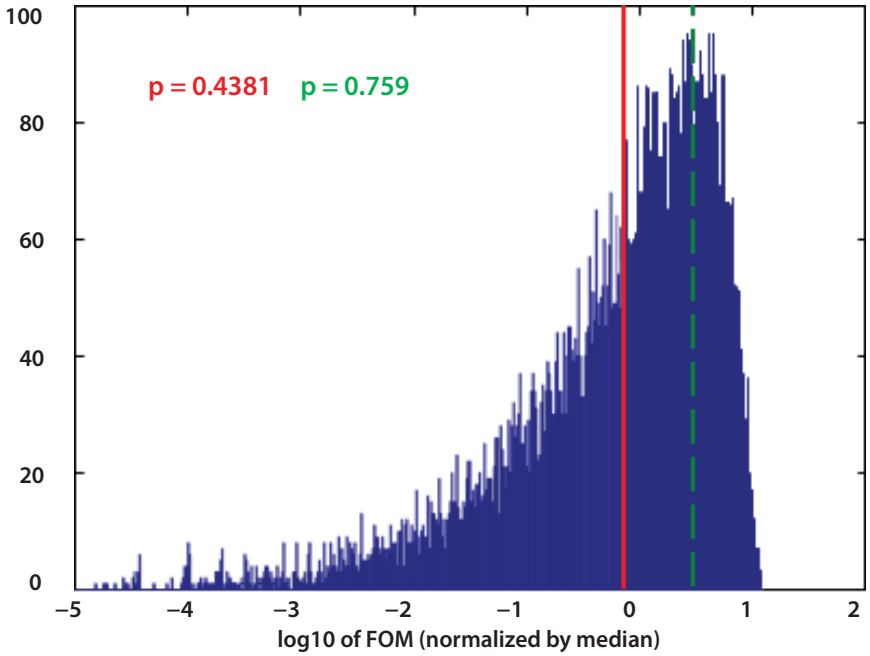

Figure 2. Result of Analysis 1 for the participants' dataset and the control dataset compared with simulated data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM) as described in the subsection **Analysis 1**. The vertical axis denotes the counts per bin of the simulated dataset, with a total of 10,000 simulated datasets being used. The two vertical lines denote the FOM of the participants' data (red/solid) and the control dataset (green/dashed).

simulated data. As implicit in the description of this analysis in the subsection **Analysis 2**, this result means that the distribution of the 20 participants' results (regarding their individual hit rates) does not significantly deviate from the expected distribution under a null hypothesis.

The probability for the result of the control dataset to have occurred by chance (null hypothesis) is $p = 0.759$, and thus also not significant. Table 4 with the individual participant results can be found in the Appendix section ***Individual Participant Results from Analysis 1***.

Analysis 2

Figure 3 shows the results of Analysis 2. The probability for the participants' results to have occurred by chance (null hypothesis) is $p = 0.703$, which is not significant. The probability for the result of the control dataset to have occurred by chance (null hypothesis) is $p = 0.512$, and thus also not significant.

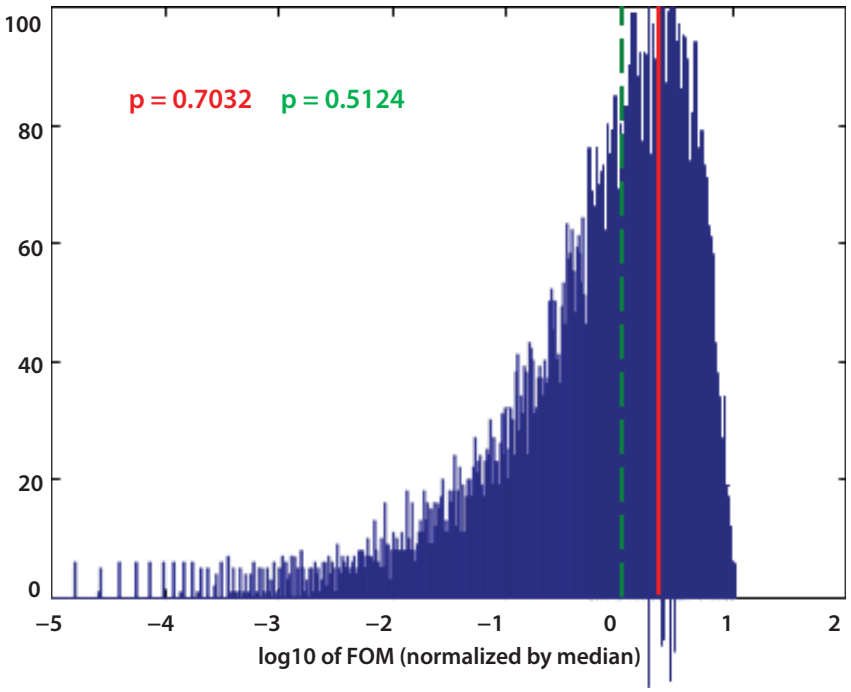

Figure 3. Result of Analysis 2 for the participants' dataset and the control dataset compared to simulated data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM) as described in the subsection **Analysis 2**. The vertical axis denotes the counts per bin of the simulated dataset, with a total of 10,000 simulated datasets being used. The two vertical lines denote the FOM of the participants' data (red/solid) and the control dataset (green/dashed).

Analysis 3

Figure 4 shows the result of Analysis 3. The probability for the participants' result to have occurred by chance (null hypothesis) is $p = 0.0949$, which is not significant using a significance threshold of $p = 0.05$. The probability for the result of the control dataset to have occurred by chance (null hypothesis) is $p = 0.983$, and thus also not significant, given that a one-sided probability had been specified. The observation that the control data are located on the right side of the distribution led to the post hoc analysis described in the next section.

As an additional illustration of the result of Analysis 3, we show here the two correlation matrices for the participants (Table 1) and control data (Table 2), respectively. For the 6 psychological and 5 physical variables as described in the subsection **Analysis 3** in the section ***Pre-Planned Data***

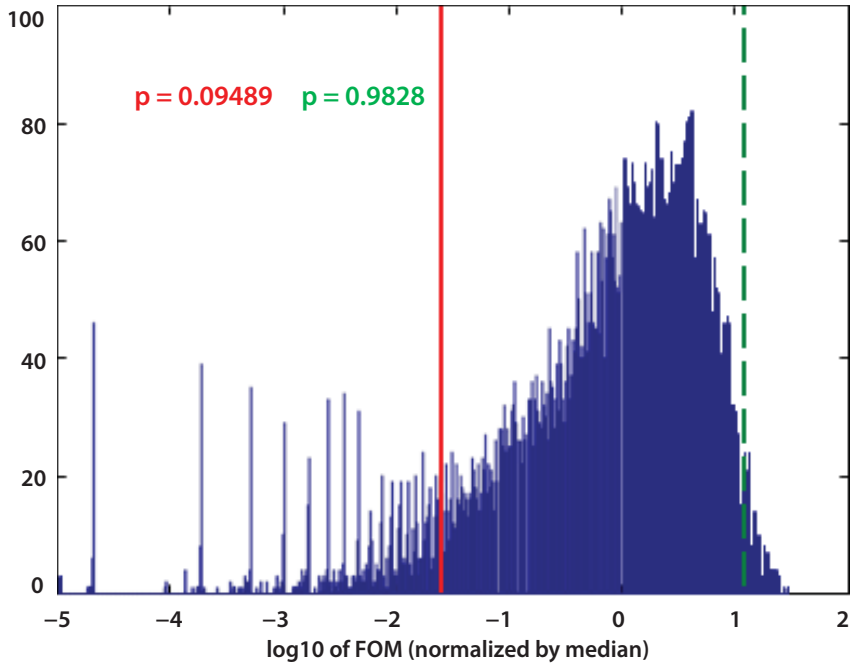

Figure 4. Result of Analysis 3 for the participants' dataset and the control dataset compared with simulated data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM) as described in the subsection **Analysis 3**. The vertical axis denotes the counts per bin of the simulated dataset, with a total of 10,000 simulated datasets being used. The two vertical lines denote the FOM of the participants' data (red/solid) and the control dataset (green/dashed).

Analysis, we have 30 correlation factors, which are converted to p-values here to be more illustrative.

It can be observed that Table 1 contains two significant correaltions with $p < 0.05$. Another element (EX correlated with HIT) comes close to $p = 0.05$. On the other hand, the matrix for the control data, Table 2, shows no element with $p < 0.16$, which indicates why the control data are on the other side of the distribution of possible results, i.e. showing particularly low correlations between psychological and physical variables.

Post Hoc Analysis

Analysis reported in this section has been performed post hoc and as such does not contribute to the statistical outcome of the pre-planned analysis.

As a post hoc analysis for Analysis 3, one can combine the results for the participants' and the control data and evaluate their combined

TABLE 1
Matrix Arrangement of p-Values for the 30 Correlations of Participant Data

Participant	HIT	ACR	RUN	EXC	GNG
TAS	0.7761	0.9046	0.2075	0.1703	0.8942
SG	0.5045	0.6657	0.7270	0.3037	0.6783
SENS	0.4911	0.9697	0.7347	**0.0032**	0.4220
TRANS	0.8221	0.8719	0.4940	**0.0285**	0.6046
EX	0.0538	0.7237	0.6399	0.2527	0.1794
MED	0.4638	0.2055	0.3934	0.1236	0.8099

p-Values smaller than $p = 0.05$ are shown in bold.

TABLE 2
Matrix Arrangement of p-Values for the 30 Correlations of Control Data

Control	HIT	ACR	RUN	EXC	GNG
TAS	0.3136	0.2461	0.3898	0.1607	0.6942
SG	0.8878	0.6176	0.6335	0.8773	0.3304
SENS	0.9924	0.9899	0.6944	0.8277	0.4205
TRANS	0.8063	0.5807	0.8969	0.4840	0.3003
EX	0.4275	0.2052	0.6907	0.7598	0.8595
MED	0.6755	0.6218	0.9421	0.5488	0.9242

significance, using a one-sided probability. In this case the prediction is that the participants' data are in the direction of high correlations, and the control data in the direction of low correlations (i.e. $p = 1 - 0.983 = 0.017$ for the control data, as pertaining to the right hand side of the distribution).

The combined probability of $p = 0.0949$ and $p = 0.017$ for uniform distributions on [0,1] is $p = 0.012$. However, while this can be called significant, even if this analysis had been pre-specified as Analysis 3, when combined with the results of Analysis 1 and Analysis 2, the combined p-value would still only be $p = 0.082$.

Two types of statistical background estimation. For the pre-defined analysis, 10,000 complete sets of simulated data derived from a Mersenne twister algorithm were used (Matsumoto & Nishimura 1998). This method relies on the assumption that the generation algorithm is sufficiently random

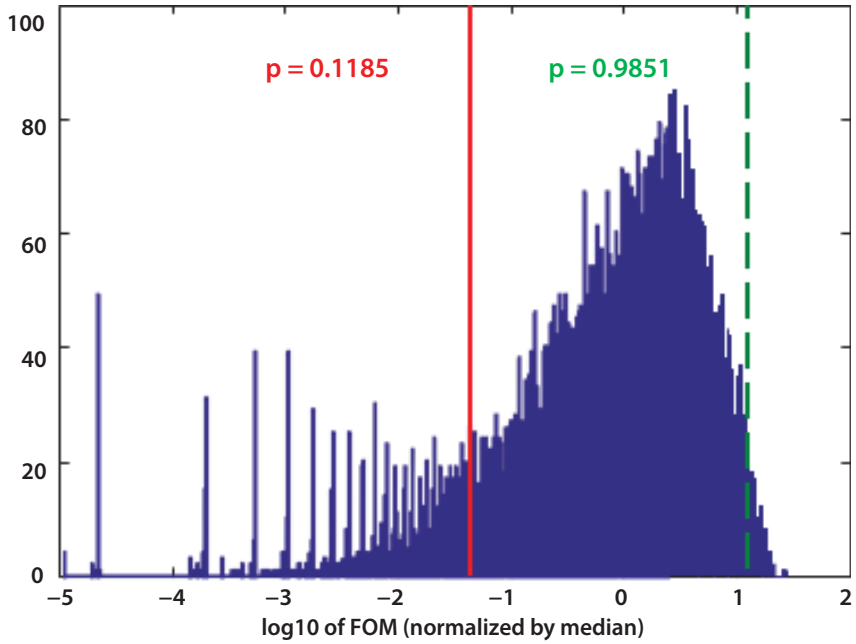

Figure 5. Result of Analysis 3 for the participants' dataset and the control dataset compared with random permutations of participant data. The two vertical lines denote the FOM of the participants' data (red/solid) and the control dataset (green/dashed) for the appropriate permutation. The background distribution is similar to the one derived from simulated data (compare with Figure 4). The combined probability of $p = 0.1185$ and $p = 0.0149$ for uniform distributions on [0,1] is $p = 0.013$.

for the purpose of the study. While data could also be generated with a hardware random number generator, the amount of required data (of order 10^{11} bits to feed the Markov chain) makes this slightly non-trivial, and a sufficiently fast hardware RNG was not at hand. Another way to estimate the background distribution is to use participants' or control data, but use many permutations of these with respect to the psychological data to which they are to be correlated. For Analysis 3 this means that the association of physical data (derived from the output of the RNG) is randomly permutated 10,000 times with respect to the psychological data. This type of background generation has been performed for Analysis 3, using participants' data.

The result of this permutation analysis is shown in Figure 5. The background distribution and the estimated probabilities are similar to the background distribution and probabilities from the simulated data in Figure 4, which corroborates the result derived from simulated data, and vice versa.

Other statistics. It may be interesting to look at the data in this experiment in a more familiar way, at the overall hit rate over all participants.

TABLE 3
Basic Statistics of Experimental Data

	N	$N_{1's}$	$N_{0's}$	z-score
Participant right intention	360,000	179,871	180,129	–0.745
Participant left intention	360,000	179,763	180,237	1.368
Control right intention	360,000	179,934	180,066	–0.381
Control left intention	360,000	179,790	180,210	1.212

N denotes total number of bits for each condition. $N_{1's}$ denotes the number of '1' bits. $N_{0's}$ denotes the number of '0' bits. z-Values have been calculated with Equation (1) and Equation (2).

Table 3 shows basic statistics for the participants' data and control data split by left and right intention. This is the classical way of analyzing data from this type of experiment. No z-score is significant for any of the 4 datasets. It may look slightly surprising that all numbers of $N_{1's}$ and $N_{0's}$ are in the same direction. However, the author would attribute this to chance, since the evaluation of the random event generator yielded no deviation from chance expectation, as described in the Appendix section ***The Binary Random Number Generator***.

Discussion

The result of Analysis 1 does not confirm the hypothesis that the distribution of individual results from the 20 participants would deviate significantly from the expected distribution under a null hypothesis. Even though the number of participants has been smaller by a factor of 2, compared with the study in Grote (2015), it seems that there is no hint of an anomalous distribution.

The results of Analysis 2 can hardly be further commented on. This analysis was exploratory in the sense of the hypotheses put forward. However, the analysis was strictly pre-specified.

The result of Analysis 3 is more interesting. Even though the main outcome is not significant with $p = 0.0949$, it is notable that the control data are located toward the right side of the distribution of the simulated data (see Figure 4). This means that the control data show significantly less correlation (between psychological and physical variables) than to be expected given the simulated data. While this could be interpreted as a

chance fluctuation, it is at least noteworthy that von Lucadou proposed that the control data might be part of the *operational closure of the system*, and thus be part of the experiment as a whole. Von Lucadou and others have used the *difference* between experimental data and control data to estimate the overall significance of experiments (von Lucadou 1986, 2006, Walach et al. 2016), thus including the control data in the analysis. However, Walach et al. (2016) find that their control data mostly conform to expectation values under a null hypothesis. In the study here, the author chose to use only the experimental data in comparison with the simulated data as a result of Analysis 3. This decision was made since it seemed more plausible to this author that an effect (if existent at all) would more likely show up in the main experimental data and not in the control data. Perhaps the fact that the control data of Analysis 3 is significantly shifted toward the low-correlated side of the distribution is yet another Trickster manifestation?

Similar to postulating a Trickster effect would be to speculate on experimenter-psi as a source of the observed result. See Parker and Millar (2014) for a more recent overview of experimenter-psi. It is interesting to note that in Analysis 3, a significant result only can be obtained by correlations across participants. There is no way an individual participant can 'score high' in this type of analysis, since each participant is only evaluated as part of an ensemble of participants. This fact may (or may not) make this type of analysis more prone to experimenter-psi.

We can note that the post hoc analysis using the difference between experimental and control data in Analysis 3 yields a probability to have occurred by chance of $p = 0.012$ under a null hypothesis. However, even when this type of analysis would have been pre-specified for Analysis 3, the combination of Analysis 1, Analysis 2, and Analysis 3 still only would yield a combined p-value of $p = 0.082$.

For future replications of correlation matrix experiments, the number of participants to be employed seems an open question for this author. While the study in Walach et al. (2016) employed about 300 participants, the study here employed 20. However the p-values for both experiments are comparable, of order $p = 0.01$ when looking for the difference between participant data and control data, and using simulated or permutated data to estimate the background. Based on this finding, one may wonder whether the number of participants plays an important role. Perhaps a useful measure for this kind of experiment could be the total interaction time between humans and the machine, which is different for the 2 experiments, but less so than the number of participants: The total interaction time for the CMM experiment in Walach et al. (2016) was 125 h while it was 20 h for the CMM experiment reported here.

Acknowledgments

The author thanks Walter von Lucadou and Eberhard Bauer for fruitful discussion and comments. This study was performed as private research by the author.

Notes

[1] A *Schmidt* process (terminology used by W. von Lucadou) is a process where generated random bits are statistically independent events. In a Markov process, the actual random event has a non-zero statistical dependence on the last internal state of the Markov process. This is further detailed in the Appendix section ***The Binary Random Number Generator***.

[2] The check-sums were generated by the device from the sum of all transmitted bytes modulo 256. The device would transmit this check-sum after a block of data had been transmitted, and the receiving computer compared this check-sum with the one it calculated from the received data.

[3] The author is aware of possible criticism of p-values for some domains of research and hypothesis testing. However, p-values as used in classical (frequentist) statistical analysis still have their merits and reasonable domains of application, as pointed out by an overview article on Bayesian and classical hypothesis testing (Kennedy 2014).

[4] Of course, in principle it may be possible to calculate the likelihood of the employed tests analytically; however, a Monte-Carlo approach was chosen here for simplicity and for better transparency of the data analysis. Further, the Monte-Carlo method makes it straightforward to combine different statistical tests and analyses that may be overlapping. The analytic approach would be exceedingly complex in this case. However, care has to be taken to assure that the random number generator used for the background distribution suffices for the intended usage. For the case here, different algorithms have been compared with no significant differences found in the resulting distributions relevant for this analysis. Another approach is to use the existing dataset with random incursion points (i.e. random permutations of the data) to generate the background distribution. This was performed for Analysis 3 and is described in the subsection **Two types of statistical background estimation**.

[5] See the subsection **A strange anecdotal occurrence: Trickster at play?**, though, for an anecdote about this testing.

[6] Equation 2 is an approximation. However, since simulated data with the same statistic as the experimental data are used to estimate the background, the exact statistic used does not matter. Just counting the number of hits

for each participant would thus yield the same result for this analysis.

[7] If both values are equal, an independent random bit is generated from the hardware random generator to resolve the tie. For consistency, the same Markov algorithm is used to generate the Monte Carlo data for the background distribution.

[8] Upon suggestion of the current author, this type of analysis has been incorporated in the most recent replication of the CMM experiment, as reported in Walach et al. (2016).

References Cited

Atmanspacher, H., Römer, H., & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics*, *32*:379–406.

Bierman, D. J., Spottiswoode, J. P., & Bijl, A. (2016). Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology. *PLOS One*, *11*(5):e0153049.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators—A meta-analysis. *Psychological Bulletin*, *132*(4):497523.

Filk, T., & Römer, H. (2011). Generalized quantum theory: Overview and latest developments. *Axiomathes*, *21*: 211–230. arXiv:1202.1659

Fisher, R. A. (1970). *Statistical Methods for Research Workers*, 14th edition. Edinburgh/London: Oliver & Boyd.

Grote, H. (2015). A correlation study between human intention and the output of a binary random event generator. *Journal of Scientific Exploration*, *29*(2):265–290.

Hansen, G. P. (2001). *The Trickster and the Paranormal*. Xlibris. ISBN 978-1-40100-082-0.

Jahn, R. G., & Dunne, B. J. (1986). On the quantum mechanics of consciousness, with application to anomalous phenomena. *Foundations of Physics*, *16*(8).

Kennedy, J. E. (2003). The capricious, actively evasive, unsustainable nature of psi: A summary and hypothesis. *The Journal of Parapsychology*, *67*:53–74.

Kennedy, J. E. (2014). Bayesian and classical hypothesis testing: Practical differences for a controversial area of research. *Journal of Parapsychology*, *78*(2):170–182.

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, *55*:141–163.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, *8*(1):330.

Odling-Smee, L. (2007). The lab that asked the wrong questions. *Nature*, *446*(March):10–11.

Pallikari, F. (2015). Investigating the nature of intangible brain–machine interaction. *Journal of Social Sciences and Humanities*, *1*(5).

Parker, A., & Millar, B. (2014). Revealing Psi Secrets: Successful experimenters seem to succeed by using their own psi. *Journal of Parapsychology*, *78*:39–55.

Radin, D., & Nelson, R. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics*, *19*(12):1499–1514.

Radin, D., & Nelson, R. (2003). Meta-analysis of mind–matter interaction experiments: 1959–2000. In *Healing, Intention and Energy Medicine* edited by W. Jonas & C. Crawford, London: Harcourt Health Sciences.

Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on Bösch, Steinkamp, and Boller (2006). *Psychological Bulletin, 132*(4):529–532.

von Lucadou, W. (1986). *Experimentelle Untersuchungen zur Beeinflussbarkeit von stochastischen quantenphysikalischen Systemen durch den Beobachter*. Frankfurt: H.-A. Herchen Verlag.

von Lucadou, W. (2006). Self-organization of temporal structures—A possible solution for the intervention problem. In *Frontiers of Time. Retrocausation—Experiment and Theory* edited by Daniel P. Sheehan, *AIP Conference Proceedings 863*, Melville, NY: American Institute of Physics, pp. 293–315.

von Lucadou, W., Römer, H., & Walach, H. (2007). Synchronistic phenomena as entanglement correlations in generalized quantum theory. *Journal of Consciousness Studies*, *14*(4):50–74.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. J. (2015). A skeptical eye on psi. In *Extrasensory Perception: Support, Skepticism, and Science* edited by E. May and S. Marwaha, New York: Praeger, pp. 153–176.

Walach, H., von Lucadou, W., & Römer, H. (2014). Parapsychological phenomena examples of generalized nonlocal correlations: A theoretical framework. *Journal of Scientific Exploration*, *28*(4):605–631.

Walach, H., Horan, M., Hinterberger, T., & von Lucadou, W. (2016). Evidence for a generalised type of nonlocal correlations between systems using human intention and random event generators. *PLOS One*.

Appendix

The Binary Random Number Generator

The random number generator (RNG) is a hardware RNG. Figure 6 shows a simplified schematic of the RNG components. The hardware RNG is based on the differential thermal noise of two resistors. The difference of the resistors thermal noise voltage is amplified and fed to the input of a comparator, comparing the noise voltage to its time average. This yields a random sequence of logic high and low levels at the output of the comparator with close to equal distribution, but which is still sensitive, for example, to offset voltage drifts of the involved amplifiers, etc. Therefore, in order to better equalize the distribution of the data, the bit stream is fed to a frequency divider which toggles its logical output on the transitions from high to low of the comparator output. This corresponds to a frequency division by a factor of two, and is a technique to equalize over time the high- to low-level ratio of a binary signal. On average, the divider registers 65 high-to-low transitions of the comparator per millisecond, corresponding to an average count frequency of 65 kHz.

This stream of randomly alternating logic high/low levels is fed to a microcontroller that controls the whole experiment. Within the microcontroller, the random bit stream from the hardware generator is sampled at a frequency of 200 Hz and fed to a 16-bit long shift register at this frequency, such that every 5 ms a new random bit is fed into the shift register.

To generate one random bit (we call this bit *b*) for the main experiment (i.e. a bit to be 'influenced' according to the participants' intentions), the software

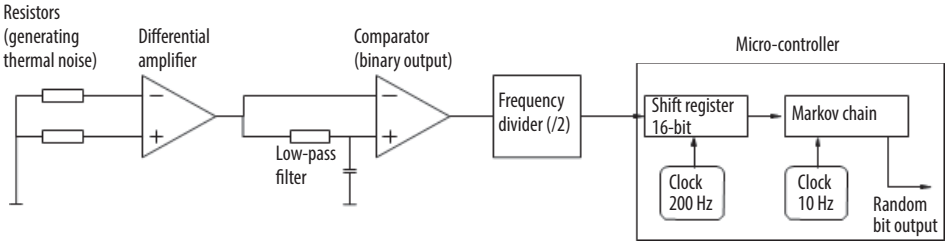

Figure 6. Schematic of the binary random number generator. See text for description.

of the microcontroller performs the following operations:

- The shift register is read out obtaining a 16-bit long word (we call it *A*). This word *A* is formed by the 16 random bits that have been fed into the shift register during the last 16 ∗ 5 ms = 80 ms. Since the individual 16 bits of *A* are random, *A* has a uniform distribution on the interval of integer numbers [0.65535].
- Word *A* is then compared to the 16-bit word that had been obtained in the previous sampling of the shift register (we call it *A*−1). If the actual word is larger than the previous one ($A > A-1$), a logical 1 is the output bit, such that $b = 1$. If it is smaller ($A < A-1$), the output bit is a logical 0, respectively, such that $b = 0$.[7] This procedure constitutes a 1-step Markov chain.
- In the last step, the value of word *A* is assigned to word *A*−1 to be used in the next iteration of these steps.

This procedure is executed 10 times per second, and thus, for the purpose of the main experiment, random bits *b* are generated with a rate of 10 Hz.

In the following, the bits "1" will be referred to as the "high bits" whereas the "0" bits will be referred to as the "low bits". A test run of this RNG comprising $N = 57{,}565{,}280$ (57 million) bits yielded $n_h = 324 + N/2$ high bits, corresponding to 50.0000056% of the cases. The corresponding z-value is $z = 0.148$, as calculated with Equations (1) and (2) above in the **Analysis 1** subsection of the ***Pre-Planned Data Analysis*** section.

The functioning of the hardware RNG was monitored automatically throughout the experiment. This monitoring was done by counting the number of high to low transitions of the random noise generator for each second, and requiring that a threshold number of transitions was passed. No error on the hardware RNG occurred during the regular experimental time of the participants.

Notes on the Correlation Matrix Method

A correlation matrix, as introduced by von Lucadou (1986), is simply the arrangement of all calculated correlation factors (or their respective p-values) in the form of a matrix, for the purpose of illustration. However, there are two questions arising about how to evaluate the matrix elements (i.e. the correlation factors) with respect to their combined statistical significance.

First, we need a method of how to combine the matrix elements into one figure of merit or combined statistic. The chosen method here, for the 30 correlation results, takes each correlation factor into account, forming one quantitative outcome of all matrix elements combined, as described above. In contrast, the method used by Von Lucadou uses only those correlation factors that are above a threshold value, and counts their number of occurrences as the combined statistic. Both methods are similar in principle, but here the first method was chosen on the hypothesis it would be more suitable for a small number of total correlations, and may also be more sensitive altogether, since no matrix elements are omitted from the analysis.

Secondly, after we have established a combined figure of merit of all matrix elements, we need to assess the statistical significance of this figure of merit (the participants' result) against an expectation value or against the control data. Due to the fact that at least the psychological variables, but perhaps also the physical variables, can be expected to correlate among each other, a comparison of the participants' data with a large set of simulated (Monte Carlo) data (i.e. the *correlations* of the simulated data with the participants' psychological data), or with a set of random permutations among psychological and physical data, seems the only way to establish a valid background distribution for this kind of analysis.[8]

Individual Participant Results from Analysis 1

TABLE 4
z-Scores of the 20 Participants for Analysis 1, Ranked by z-Score Value

This score (as defined in the subsection **Analysis 1**) is a measure of how well the participants succeeded in 'influencing' the galvanometer needle in the desired direction. Also shown are the expectation values for the z-scores.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
z-Score	1.80	1.55	1.53	1.35	1.26	0.95	0.79	0.59	0.35	0.16
Exp.value	1.87	1.41	1.14	0.92	0.75	0.59	0.45	0.31	0.19	0.06
Rank	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
z-Score	0.16	0.16	–0.02	–0.24	–0.27	–0.44	–1.24	–1.59	–1.95	–2.94
Exp.value	–0.06	–0.19	–0.31	–0.45	–0.59	–0.75	–0.92	–1.14	–1.41	–1.87