Journal of Scientific Exploration, Vol. 29, No. 2, pp. 265–290, 2015 0892-3310/15

RESEARCH ARTICLE

A Correlation Study between Human Intention and the Output of a Binary Random Event Generator

H. Grote

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

Submitted May 1, 2014; Accepted March 17, 2015, Published June 15, 2015

Abstract—This paper reports on a correlation study between human intention and the output of a binary random number generator. The study comprises a total of 288 million bits from 40 equal sessions, each on a different human participant. Each participant spent 2 hours of time 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-hand side of its current position, according to the instantaneous output of the random number generator. The data analysis procedure was defined before looking at the data. Out of four pre-defined analyses, one was found to be significant with a probability $p = 0.0366$ that this result occurred by chance under a null hypothesis. The combined analysis of the four individual analyses is found to be not significant, with $p = 0.2655$ 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 not to be in doubt—see for example Radin and Nelson (1989, 2003) and Jahn and 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 interpretation of the research results, which makes it difficult for new researchers to form an opinion on the research performed to date, as visibly exemplified in the dispute on the interpretation and validity of meta-analysis of existing mind–matter experiments (Bösch, Steinkamp, & Boller 2006,

Radin, Nelson, Dobyns, & Houtkooper 2006). 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 (that is 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 explanation of a putative correlation effect, such as the interpretation as entanglement correlations in a Generalized Quantum Theory (Atmanspacher, Roemer, & Walach 2002, Filk & Römer 2011), do exist (von Lucadou, Römer, & Walach 2007, Walach, von Lucadou, & Römer 2014).

For these reasons, it seems of value to the field if new mind–matter experiments are performed from time to time, in particular if new researchers conduct such experiments and possibly introduce new aspects to the experimental approach. They should also serve to avoid 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 in Kennedy (2003), von Lucadou, Römer, and Walach (2007), and Walach, von Lucadou, and Römer (2014) and references therein.

The primary intent of the study described in this paper was not that of investigating a specific aspect of putative mind–matter correlation, but rather to contribute with an original new experiment to this field of research. However, beyond the standard analysis of looking for correlations between the output of the binary random number generator in the direction of the participants' intention (see **Analysis 1**), more complex types of data analysis were performed in this study, which were partially inspired by the correlation matrix technique that has been used by von Lucadou and others (von Lucadou 2006, von Lucadou, Römer, & Walach 2007).

One of the basic ideas of this technique is to not make predictions about deviations of any particular statistical test of the data, but rather to look at the number of deviations of a total ensemble of statistical texts. For this purpose, a combined figure of merit of a number of statistical tests is defined, and compared with its corresponding expectation value. This is further detailed in the section **The Data Analysis Procedure**. The analysis of data was defined before any of the data was actually analyzed, and it was decided 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 **The Experiment**, the experimental setup is described, followed by **The Data Analysis Procedure** on the predefined data analysis plan. The results of the analysis are presented in the **Results** section. Finally, the last section, **Discussion**, contains a brief discussion of the analysis and results in the context of existing research and terminology in the field.

The Experiment

The experiment described in this paper was designed and conducted by the author. Participants were 40 people (including the author) with different relationships to the author (friends, friends of friends, work colleagues, etc.) who were interested in the topic, and willing to spend two hours each in actual experimentation time. With one exception, none of the participants had ever taken part in any similar experiment of this kind. The participants' ages spanned from 15 to 73 years old, and participants included both genders.

Each participant had agreed to carry out 120 runs, with each run lasting 60 seconds. A single run would always begin with the participant selecting whether he/she would try to influence the motion of the needle of a galvanometer display to the left-hand side or to the right-hand side during that run. Then the participant would press the start button to begin the 60-s run. While the run was active, a red light was lit in the background of the display needle, to signal the participant that the run was going on.

During each 60-s–long run, random binary events would be generated at a rate of 1,000 per second. The draw of a logical 0 would result in the step of the display needle to the left-hand side of its current position, while a logical 1 would result in a step of the needle to the right-hand side of its current position. In this way, 60,000 binary random draws were accumulated during each 60-s run, resulting in a random walk of the needle.

Figure 1 shows an image of the experimental device in active display.

Figure 1. The experimental device, photographed in a state with active display, simulating a real data-taking run. The needle (upper right) has moved to the right-hand side during this run, as a result of the random walk, accumulating the binary random generator output. The alphanumeric display on the upper left side of the image shows the name of the participant (here Test), the chosen intention (here in German *Rechts* for right), the actual accumulated random generator bits as deviation from equal distribution (here 97), and the remaining seconds for the actual run (30).

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 preferably 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 60 s could not be interrupted by any means, but the participants were free to distribute the time to perform the runs at their choice of time. The participants could choose for any run between left or right intention, but had to respect the constraint that out of the 120 runs left and right intention had to be picked the same number of times, 60, 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 2 hours of data in total. Each participant committed to collecting these 2 hours 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 2 months. The experimental data-taking started in the summer of 2009 and concluded late in 2012, when the number of 40 participants had been reached. Up to 4 participants could share the device (e.g., members of a family) by freely distributing experimentation time among them. 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 experiment data was stored in two different formats in the device, to be safe against errors in the storage. No such error occurred. Data before storage was reduced to 4,000 cumulated bits each, corresponding to 4 s of data. This reduced dataset was used for the analysis of the experiment. The data was transmitted to a personal computer after 1–4 participants had completed their runs, and the device was prepared for the next participant(s). The data transmission to the personal computer used checksums to be safe against transmission errors, and no such errors occurred. In addition to the participants' actual data, 2 different sets of control data were taken, which were not explicitly subject to any interaction with the intention of any participant:

Control set a): Whenever a participant decided to end a series of runs (but at the latest after 30 consecutive runs), the device automatically collected data from the random number generator without feedback to the display. During these times, the message ***Kontrolllauf*** (German for control run) was displayed in the alphanumeric display of the device, and no particular instruction was given to the participants during these times. This way, reference data of the same length for left and right intentions was taken, that is 1 hour of data for each participant.

Control set b): Between participants (that is 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 was automatically generated. For this purpose, dummy persons with names 01 to 40 were generated by the conductor, and when the device recognized a dummy participant name 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 but satisfied the required equal total number of left and right intentions as for the real runs. This way a complete set of 40 dummy participants was created and spread throughout the years of acquisition of participants' data, which will be 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 times and in environments of their choice. While this may appear as 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 be more involved in their effort to influence the needle. Ultimately, even in the laboratory, the conductor of the experiment has no control over whether the participant asserts influence on the device according to the prestated 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 also carries an element of being different.

The Binary Random Number Generator

The random number generator (RNG) is a hardware RNG combined with a software RNG. **Figure 2** shows a simplified schematic of the RNG components. The hardware RNG is based on the differential thermal noise of two resistors. The difference in the resistors' thermal noise voltage is amplified and fed into 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. The data gets better equalized in distribution by feeding it into a frequency divider, which toggles its logical output on the transitions from high to low of the comparator output. This corresponds to a frequency divi-

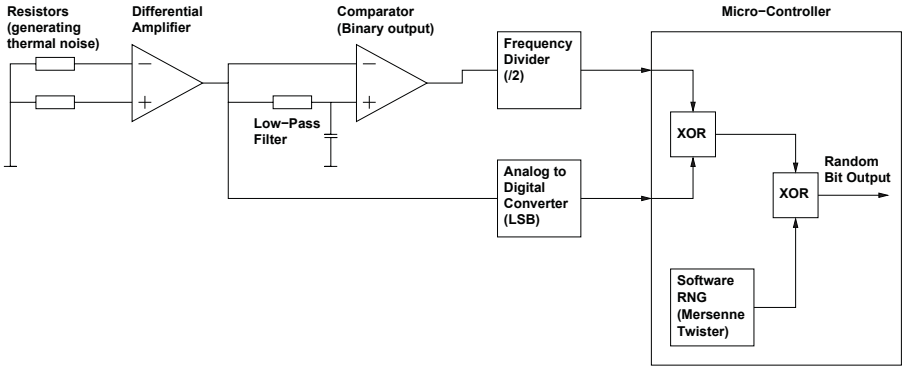

Figure 2. Schematic of the binary random number generator. See text for description.

sion by a factor of two, and is a technique to equalize in time the high-level to low-level ratio of a binary signal. One random bit is then generated by reading the logical output state of the frequency divider. On average, the divider registers 65 high-to-low transitions of the comparator per millisecond, corresponding to an average count frequency of 65 kHz. The quality of randomness of this bit is further increased by a logical exclusive–or operation with the least significant bit (LSB) of one sample of an analog-to-digital converter, which samples the actual noise voltage at the time the random bit is requested. This resulting bit is the output of the hardware random number generator.

In order to further improve the quality of randomness, and to get a very high level of security against potential (albeit unexpected) misbehavior of the RNG hardware as well as principal fraud attempts by the participants, it was planned to combine the output of the hardware RNG with the output of a software RNG. If the outputs of two randon number generators are combined, (e.g., with an exclusive–or operation in the case of single bits), the resulting output is of a higher quality of randomness, as long as the two generators are uncorrelated. There seems to be no reasonable doubt that the latter is the case when a hardware generator is combined with a software generator.

While some researchers may assume that this design (the combination of a hardware RNG with a software RNG) will make any sort of influence of the RNG harder or impossible, there may be evidence in the literature that this is not the case, and that significant results may be obtained with substantially different sources of randomness: See for example Schmidt

(1987) for a discussion on the use of different types of random number generators in MMI experiments.

For the software RNG here, an algorithm called Mersenne twister (Matsumoto & Nishimura 1998) was chosen, with a simplified implementation named TT800. This algorithm has a period length of $2^{800} - 1$ and was seeded with numbers from the hardware RNG prior to the start of a participant's contribution.

The combined random number generator used in the experiment uses the exclusive–or operation to combine the subsequent outputs of both generators: For each bit obtained from the hardware RNG, a bit from the software RNG is generated and X–or'd with the hardware RNG bit. The resulting output is used as a random bit for the main experiment (see also **Figure 2**).

As noted above, for the purpose of the main experiment, random bits are generated with a rate of 1 kHz, thus producting 1,000 bits per second. 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 combined RNG comprising N = 9508571000 (9.5 billion) bits yielded n_h = 4.754282524 billion high bits, corresponding to 49.9999687% of the cases. The corresponding z value for the null hypothesis (no bias) is

$$z := \frac{n_h - E[n_h]}{\sigma[n_h]} = -0.611 \qquad (1)$$

where $E[n_h]$ and $\sigma[n_h]$ are the expected value and the standard deviation for n_h, respectively, in the absence of bias. $E[n_h] = Np_{nb}$ and $\sigma^2[n_h] = Np_{nb}(1-p_{nb})$ with $P_{nb} = 0.5$ being the hit probability of a single trial. The corresponding cumulative chance probability (similar as defined in **Equation 4**) is 47.6%.

A test run of the hardware RNG comprising N = 9508571000 (9.5 billion) bits yielded high bits in 49.99866% of the trials. The corresponding z value for the null hypothesis (no bias) is $z = -2.6$. This excess of low bits has a chance probability to occur in a realization of this same experiment that is less than 5 per thousand (cumulative chance probability of 0.45%), which indicates that the hardware RNG is not free from bias. This small bias is not relevant (but reported for completeness), since it is subsequently removed by the combination with the software RNG.[1]

A test run with 9,508,571,000 (9.5 billion) bits of this software RNG yielded high bits for 50.00000371% of all bits generated, corresponding to a z value of $z = 0.072$ and a cumulative chance probability of 50.2%.

While the combination of the hardware RNG with a software RNG

already serves as a safeguard against a possible (in principle) malfunctioning of the hardware RNG or fraud attempt, the functioning of the hardware RNG was also automatically monitored 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 be passed. No error on the hardware RNG occurred during regular experimental times of the participants.[2]

Personal Statements of Participants

To illustrate involvement and subjective experience, three of the participants have been asked to describe their perception of participating in the experiment. Here are their statements (translated from German to English by the author).

Participant S.R.: My approach to the experiment felt ambiguous. On the one hand, "This is not possible. This cannot work," which probably resembles the mainstream view around me. Also it is somewhat important to me to cling to (putative) logical reasoning, after all I am also culturally imprinted by my scientific study (of medicine), etc. On the other hand, there is the fun of resistance against all this, against this kind of all-too-fixed worldview. On performing the experiment, this kind of resistance attitude kept me going. Secondly, I was in a kind of aroused state, to enforce my will against this stupid machine. Such boosts of motivation were interrupted by phases of frustration and feelings of uselessness, in particular when I had the impression to have had a lot of failures. Altogether though—and against my expectation—I felt relatively motivated during this long experimental time.

Participant D.U.: I approached the experiment in a kind of unbiased, playful way. And like every good player, I want not only fun, I also want to score! Anyway, the experiment developed a certain dynamic: I tried different techniques, for example extremely relaxed, almost indifferent, leaving the needle almost without my intention. Then at other times I imposed pressure, or tried to make the "way back" for the needle harder if it was moving in the right direction. Of course there were also phases of resignation, but altogether I can say that I took up the fight.

Participant A.B.: The execution of the experiment was interesting. In principle I had expected it would be boring to concentrate on this little metal needle. However, after a short time I realized that I reacted strongly to the action of the needle. If it performed according to my wish, I not only had fun but also perceived it as my accomplishment, and this even against my rational conviction. This feeling of accomplishment got stronger and had quite an impression on me. Conversely, if I was not successful, I did not interpret failure to move the needle in the intended direction as my personal

fault, but rather I perceived the machine as a stubborn opponent. However, I felt motivated then to work for new success with more effort.

The Data Analysis Procedure

In order to avoid any bias, the data analysis procedure was defined before any of the data was actually looked at. Four different investigations (Analyses 1–4) were carried out, as described in the following subsections. The principal outcome of each of the four analyses is a number describing the probability that the obtained result would have occurred by chance under the null hypothesis, that is assuming no correlation between the data and the experimenters' intention.[3] The chance probability for the combined results of the four investigations is also given, taking into account possible overlaps in the four individual analyses.

Besides the comparatively simple and fully analytical **Analysis 1** (as defined below), there are two principles to be used for Analyses 2, 3, and 4.

The first principle is to estimate likelihoods of statistical test results by comparison with a large number of simulated data. This is, in essence, a Monte Carlo procedure used to estimate a background stochastic process. It 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 realizations of the statistical scores under the null hypothesis.[4] The simulated (Monte Carlo) data consists of 10,000 complete sets of data, each resembling data of a full study comprising 40 "participants."

The second principle of the data analysis procedure is to not make predictions about the outcome of individual statistical tests, but to combine the results of a number of tests in one figure of merit (FOM). This FOM can, for example, be the product of the estimated likelihoods of the applied statistical test results. The second principle was inspired by the correlation matrix technique used by von Lucadou and others, as mentioned in the **Introduction**. In the form used here, it mainly consists in a method to perform multiple analyses, as will be discussed in the subsection ***The Choice of Data Analysis***.

As detailed in the sections below, both principles are combined in the defined data analysis. The data are either combined over all participants, or separately analyzed for each participant. The statistical tests on the data will either be a single test (the integrated binomial distribution with respect to

TABLE 1
A Simple Overview of the Four Types of Analysis

	All Data Combined	Data Split by Participants
Single Binomial Distribution Test	Analysis 1	Analysis 2
Multiple Statistical Tests	Analysis 3	Analysis 4

participant intention), or multiple statistical tests of different kinds. **Table 1** gives an overview of the four types of analysis as defined in the subsections below. The control data set b), as defined in the section The Experiment will be subject to the same Analyses (1–4) as the main dataset. The analysis of control dataset b) is expected to show a high (that is nonsignificant) probability to have occurred by chance when compared with the reference data. Thus it is expected to corroborate the assumption that the reference dataset is sufficiently randomly distributed, as well as the control dataset b).[5]

Finally, we point out that the description of the experiment, the definition of the preplanned data analysis, as well as the analysis code and the complete experimental data was uploaded to the website opensciencefamework https://osf.io/ prior to the actual analysis of the data. Also prior to the actual analysis, the data on said website was marked as a read-only representation of the project (it cannot be modified) and can be made accessible to interested readers upon request.

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. Conversely, we define a miss to be a low bit when the participant's intention was to move the needle to the right, and to be a high 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. From **Equation 1** it is straightforward to see that the z value for n_{hits} over a total number of trials N is:

$$z = \frac{n_r - n_l}{\sqrt{N/2}} \tag{2}$$

where n_r are the high bits scored under right intention and n_l the high bits scored under left intention. These quantities will be determined by considering together the scores ($n_{r,p}$ and $n_{l,p}$) from all participants:

$$n_r = \sum_{p=1}^{40} n_{r,p} \qquad n_l = \sum_{p=1}^{40} n_{l,p} \tag{3}$$

The *z*-score is a useful quantity because it immediately provides a sense of the deviation of the results from the expectations. However, for the estimation of the actual chance probabilities associated with each result, it will be more convenient to refer back to the original binomial distributions.

The cumulative chance probability (null hypothesis) for the obtained results will be determined analytically here. The cumulative chance probability, $P_0(n_{hits})$, that is the probability of obtaining the measured number of hits, or greater, by chance is simply the integrated binomial probability:

$$P_0(n_{hits}) = \sum_{n'=n_{hits}}^{N} \left(\frac{N}{n'} \right) p_{nb}^{n'} (1 - p_{nb})^{N-n'} \tag{4}$$

Notes on Analysis 1. This is the classical way of analyzing this type of experiment. This analysis tests for a (positive) correlation between the participant's intention and the given task, that is to influence the display in the given direction and thus to increase the number of hits for each direction above chance expectation. The probability is defined as a one-sided probability. Note, however, that this analysis is still balanced between trials acquired under left intention and right intention.

Analysis 2

This analysis analyzes the data as detailed in the previous section (calculating *z*-scores for the number of obtained hits) but for each of the 40 participants separately, such that 40 *z*-scores are generated. These 40 *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 reference (null hypothesis) dataset. These *p* values are two-sided, with $p = 1$ if a data point is exactly in the middle of the distribution being compared to. The resulting 40 *p* values are combined (by summing over the inverse squares of *p* values) and the result is the FOM for this test. The chance probability for the value of this FOM is measured on the distribution for the same FOM derived from the Monte Carlo dataset. A one-sided probability will result in the FOM of the test data (or a lower one)

occurring by chance. This is the result of Analysis 2.

Notes on Analysis 2. This analysis is sensitive in particular to the distribution of results among the participants. It is also sensitive to deviations from randomness in directions opposite to a participant's intention.

Analysis 3

This analysis comprises a number of statistical tests for randomness (as listed below) of the acquired data of all participants combined. It is not predicted which of the pre-specified statistical tests would show a significant deviation from the expected distribution under a null hypothesis, but each of the test results (which are scalar numbers) is compared to the equivalent test results of a large number of reference 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 statistical test) are multiplied to yield a single figure of merit (FOM) of the acquired data. Finally, this FOM is compared with the distribution of the same FOMs of the reference data, and a one-sided (frequentist) likelihood results, that the actual FOM (or a lower one) of the data being tested would have occurred by chance. This likelihood is the result of Analysis 3.

Notes on Analysis 3. As mentioned above, the analysis chosen here has some similarity with the correlation matrix technique as described for example in von Lucadou (2006) and von Lucadou, Römer, and Walach (2007). 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 the Analysis 3 defined here, different physical variables correspond to different statistical tests of the data. The psychological variables in the experiment under report in this paper are just the left or right intention to influence in the direction of the needle display. In this case, the corresponding correlation matrix would consist of only 2 rows (left and right intentions), and n columns, if n is the number of statistical tests applied.[6]

This matrix could be given as a table in principle, but as defined above a figure of merit will be used instead to combine the obtained probability levels of all tests numerically (second principle). In the last step, the resulting figure of merit is compared to the set of computer-generated reference data (first principle). The statistical tests to be applied to the data are the following (tests that do not include a combination of right and left intention explicitly are performed on both intentions individually, as two separate tests, as the numbers in brackets denote):

* successful runs of 60 s length (2)
* sum of bits (2)
* standard deviation (2)
* skewness (2)
* kurtosis (2)
* chi square goodness of fit to expected binomial distribution (2)
* Ansari Bradley test if variance between right and left intention differs (1)
* distribution of sign permutations in 5-Tuples of data (2)
* correlation between left and right intention data (1)
* runs test for expected number of runs with same sign (2)
* runs test for expected number of runs with same slope (2)
* Fourier transform (Welch method) of the time series (2)
* sum of absolute difference between all consecutive values (2)
* chi square test of uniformity on 2,400 stretches of 60 s length (2)
* chi square test of uniformity on 4 stretches of 10 h length (2)

If a test result is not obviously a scalar, an algorithm is to be defined to calculate a scalar out of the test result.[7]

Analysis 4

This analysis is one step more complex than Analysis 2 and Analysis 3, and is a combination of the two: The data is first split according to the 40 participants. Then a number n of statistical tests (as listed below) is applied to each participant's data. Then for each out of the n tests and for each of the participants the following is performed.

Each test result is compared to (400,000) reference datasets and the resulting (two-sided, frequentist) probability p_{ki} is calcuated, describing the probability that this result (or a lower/higher one) occurred by chance. Here $k = 1..40$ is the number of the participant and $i = 1..n$ is the number of the applied statistical test.

From these p_{ki} values, an FOM is computed for each participant by multiplying all the p_{ki} values with $i = 1..n$ of the respective participant $k = 1..40$. The resulting 40 FOMs of the participants are then sorted and compared to reference data in a way that the highest of the 40 participants' results is ranked against the highest of all the reference data results, where the highest (of the reference data results) refers to all the highest of the (40) participants' each, of the reference dataset. In the same way, all the second-highest of the (40) participants' results are ranked against the second-highest (out of 40 each) of the reference data results. And so forth for the remaining 38 results.[8]

A final FOM is then computed by combining the participants' FOMs (by summing over the inverse squares of the p values). This final FOM is

then compared to the same FOMs of the reference data and a final one-sided likelihood will result, describing the likelihood that this result (or a lower one) occurred by chance.

Notes on Analysis 4. This analysis should be particularly sensitive to variations between individual participants (with respect to the statistical test applied) which might (if existent) be averaged out in the other analysis.

The statistical tests to be applied to the data are the following (as for Analysis 3, tests that do not include a combination of right and left intention explicitly are performed on both intentions individually as two separate tests, indicated by the number in brackets):

* successful runs of 60 s length (2)
* sum of bits (2)
* sum of bits of first half of data (2)
* standard deviation (2)
* skewness (2)
* kurtosis (2)
* single largest (/smallest) value (2)
* chi square goodness of fit to expected binomial distribution (2)
* Ansari Bradley test if variance between right and left intention differs (1)
* distribution of sign permutations in 5-Tuples of data (2)
* correlation between left and right intention data (1)
* correlation between first and second half of data (2)
* runs test for expected number of runs with same sign (2)
* runs test for expected number of runs with same slope (2)
* Fourier transform (Welch method) of the time series (2)
* sum of absolute difference between all consecutive values
* chi square test of uniformity on 60 stretches of 60 s length (2)
* chi square test of uniformity on 4 stretches of 900 s length (2)

Results

Analysis 1 Results

Figure 3 shows the full dataset obtained by the 40 participants. The total number of hits (from left and right intention) is $n_{hits} = N/2 + 2{,}018$, that is an excess of 2,018 hits over the expected number of hits $N/2 = 144{,}000{,}000$. The probability for the result of the participants' data to have occurred by chance (under the null hypothesis) is $p = 0.406$, and thus not significant.

Figure 4 shows the control dataset b), which was obtained with the experimental device running unattended for 40 dummy "participants," as described in the section **The Experiment**. The probability for the result of the control dataset b) to have occurred by chance (null hypothesis) is $p = 0.599$, which is not significant.

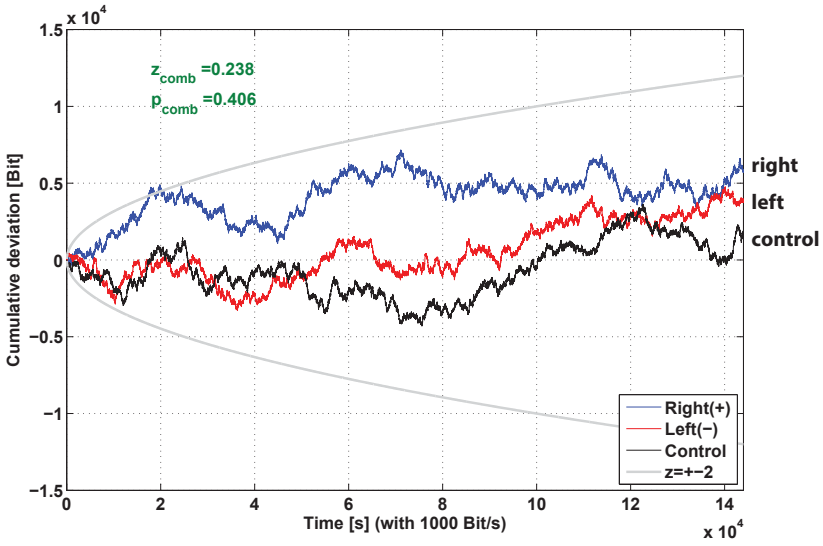

Figure 3. The full dataset obtained by the 40 participants. The horizontal axis denotes the time over which data was acquired, equivalent to the accumulated number of bits generated. The vertical axis shows the cumulated deviation from the expectation value, separated for bits obtained under right and left intention. Also shown is the control dataset a), as defined in the section **Analysis 2**, which is, however, not subject to any analysis. The grey (smooth) line denotes the level of two standard deviations. The combined probability p_{comb} of the data under right and left intention is given as defined in the section **Analysis 1**.

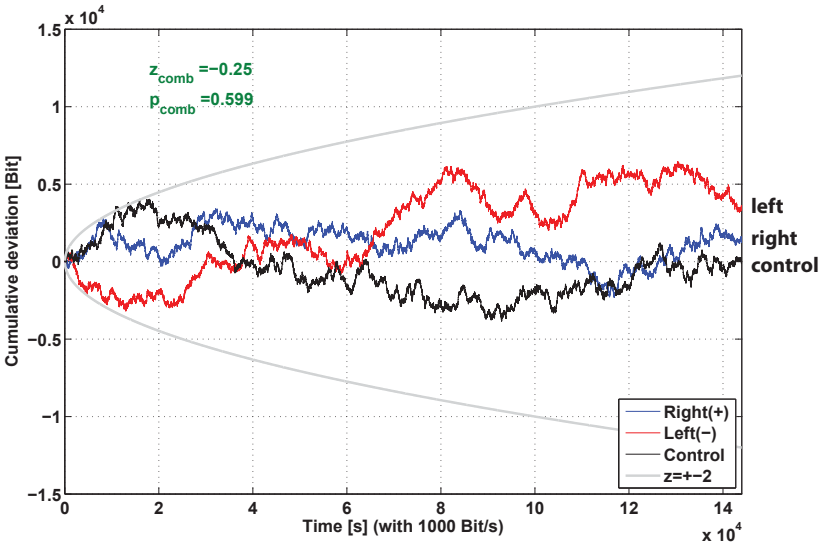

Figure 4. Control dataset b), in identical representation as the data in Figure 3.

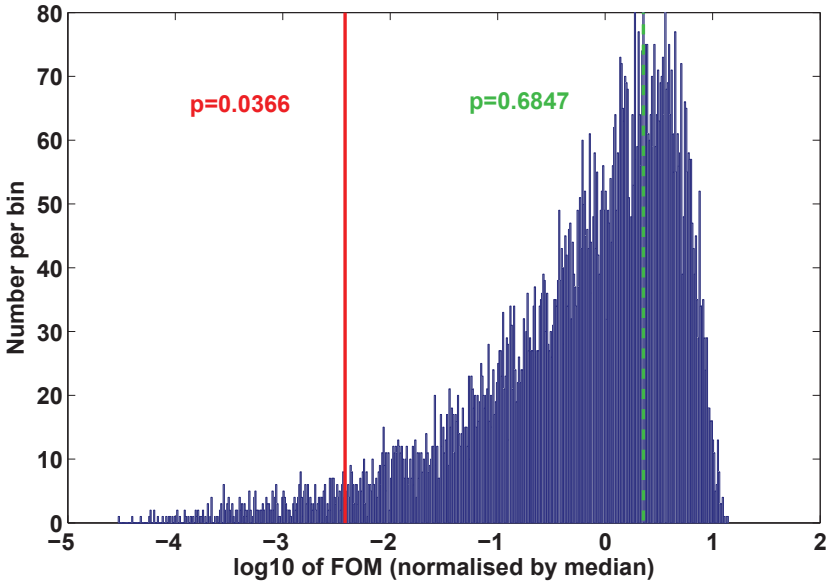

Figure 5. Result of Analysis 2 for the participants' dataset and the control dataset b) compared to Monte Carlo data. The horizontal axis denotes a logarithmic representation of the figure of merit (FOM) as described in the section **Analysis 2**. The vertical axis denotes the counts per bin of the Monte Carlo 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 reference dataset b) (green/dashed).

Analysis 2 Results

Figure 5 shows the result of Analysis 2. The probability of the participants' results to have occurred by chance (null hypothesis) is $p = 0.0366$, which is significant with respect to a 5% significance level. This probability is obtained by the fraction of more extreme results (more negative FOM) divided by the number of all results of the Monte Carlo data. As implicit in the description of this analysis in the section **Analysis 2**, this result means that the distribution of the 40 participants' results deviates significantly ($p = 0.0366$) from the expected distribution.

The probability for the result of the control dataset b) to have occurred by chance (null hypothesis) is $p = 0.6847$, and thus not significant.

For further illustration of this potentially interesting result, **Figure 6** shows the distribution of the individual participant's results, from which the FOM is calculated. The largest deviation of the participants' data from the expected distribution can be seen around the highest rank numbers (those

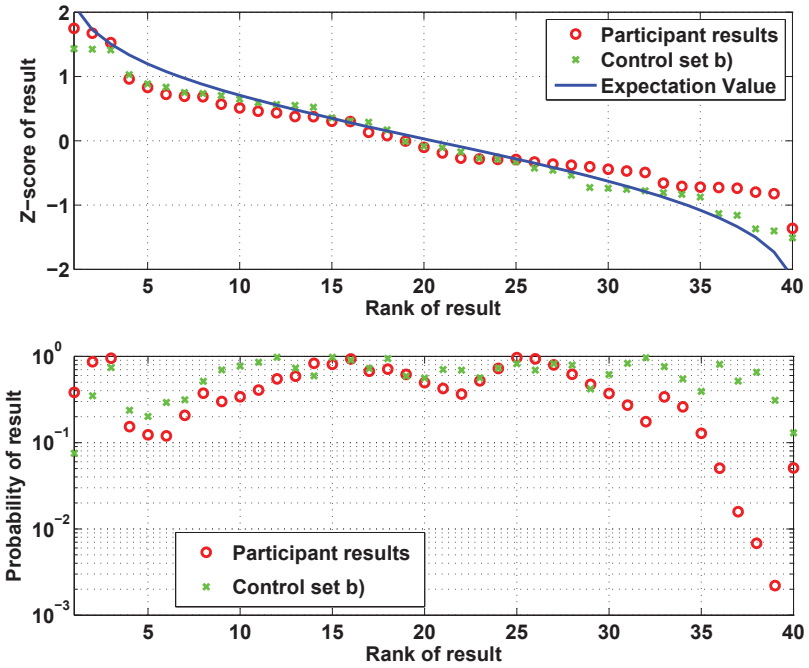

Figure 6. Individual results of Analysis 2 for the participants' dataset and the control dataset b). The horizontal axis denotes the rank (1–40) of each of 40 individual results. The vertical axis of the upper graph denotes the *z*-score of each individual result. The participants' data points are shown as (red) circles and the control data points are shown as (green) crosses. In the upper graph, the distribution of the expected *z*-scores is given as well, as the (blue) solid line, obtained from the Monte Carlo data. The lower graph shows the individual p values of the results of Analysis 2, as obtained from the Monte Carlo data. The lowest p values correspond to the largest deviations from the expected *z* values in the upper graph.

with the lowest *z*-scores in the upper graph): The ensemble of all results is slightly short of results of more negative *z*-scores. Corresponding to these deviations seen in the upper graph, the lower graph shows the *p* values of the individual results with respect to the individual expectation value of their rank.

As can been seen in the lower graph, there are 5 participants with individual results of probabilities smaller than $p \approx 0.05$. However, it would seem not quite right to isolate these individuals as extraordinary performers, since in fact no single individual in this analysis has performed significantly on his/her own (all absolute *z*-scores in the upper graph are smaller than 2). Rather it is the performance of all the participants that has to be taken into

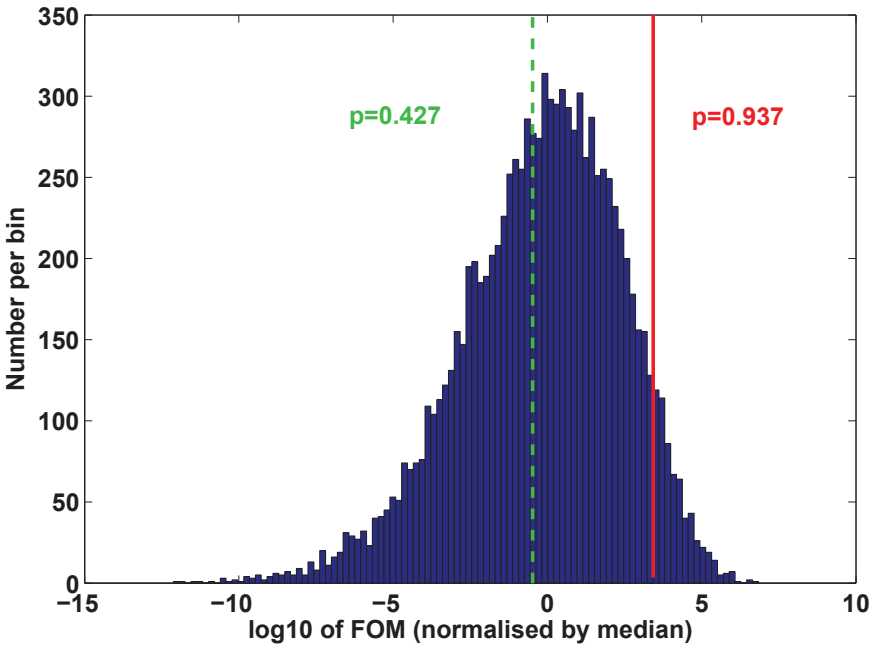

Figure 7. Result of Analysis 3 for the participants' dataset and the control dataset b) compared to Monte Carlo data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM) as described in the section **Analysis 3**. The vertical axis denotes the counts per bin of the Monte Carlo 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 reference dataset b) (green/dashed).

account for the composition of this distribution. Therefore, all participants have contributed to this result.

Analysis 3 Results

Figure 7 shows the result of Analysis 3. The probability of the participants' results to have occurred by chance (null hypothesis) is $p = 0.9411$ (one-sided, as was defined for this analysis), and thus not significant. The probability of the result of the control dataset b) to have occurred by chance (null hypothesis) is $p = 0.4468$, also not significant.

Analysis 4 Results

Figure 8 shows the result of Analysis 4. The probability of the participants' result to have occurred by chance (null hypothesis) is $p = 0.517$ and thus

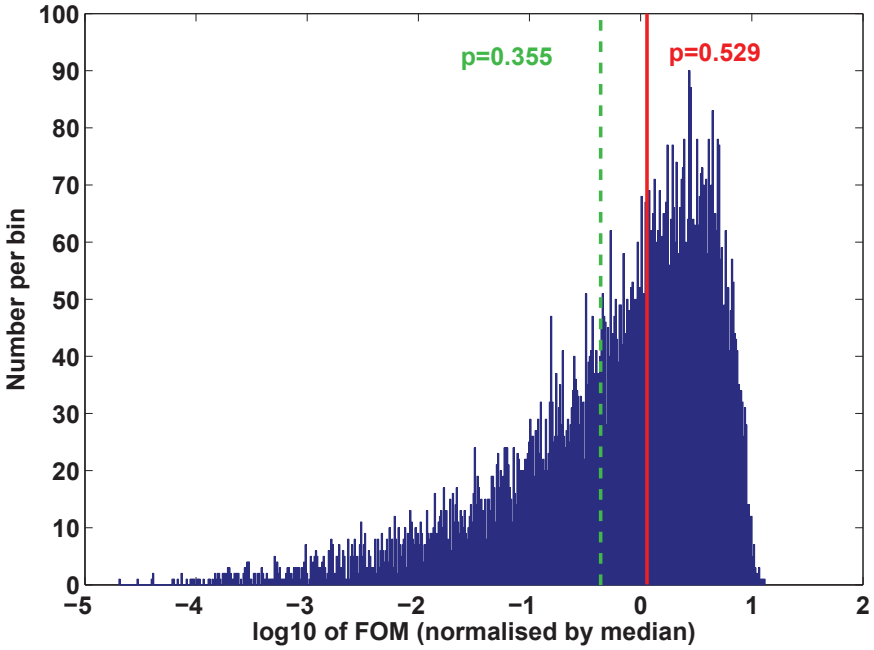

Figure 8. Result of Analysis 4 for the participants' dataset and the control dataset b) compared to Monte Carlo data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM) as described in the section **Analysis 4**. The vertical axis denotes the counts per bin of the Monte Carlo 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 reference dataset b) (green/dashed).

not significant. The probability for the result of the control data to have occurred by chance (null hypothesis) is $p = 0.374$ and thus also not significant.

Combined Analysis

The combined analysis was not predefined, but it was planned to execute a significance evaluation of Analyses 1–4 combined, in case at least one of them would be significant, or at least two would be nearly significant. The most straightforward way is chosen here, which is the calculation of a figure of merit combining the 4 results from Analyses 1–4 (by calculating the product of the 4 probability results noted in the subsections above), and comparing this FOM to the same FOM from the Monte Carlo dataset. The final result is obtained as a one-sided ranking. Because this final p value is obtained from simulated data, it includes the adjustment for possible

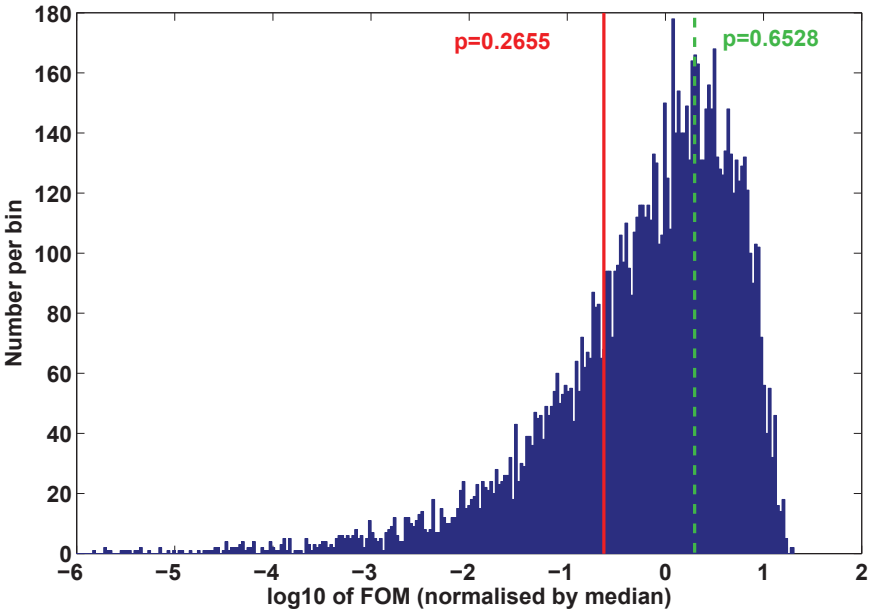

Figure 9. Result of the combined analysis for the participants' dataset and the control dataset b) compared to Monte Carlo data. The horizontal axis denotes a normalized logarithmic representation of the figure of merit (FOM). The vertical axis denotes the counts per bin of the Monte Carlo dataset, with a total of 10,000 simulated datasets being used. The two vertical lines denote the FOM of the participants' data (red/solid) and the reference dataset b) (green/dashed).

multiple analyses across the 4 predefined Analyses 1–4. **Figure 9** shows this FOM for the participants' data and the control dataset b), in comparison with the Monte Carlo data. The probability of the participants' data in the combined analysis to have occurred by chance (null hypothesis) is $p = 0.2655$, and thus not significant. The probability for the control data to have occurred by chance (null hypothesis) is $p = 0.6528$, also not significant.

Discussion

The reader may judge on the chosen analysis methods, study design, and results on his/her own; however, some discussion in the context of existing research concepts might be useful.

The Choice of Data Analysis

The basic structural description of the analysis has been given in the section **The Data Analysis Procedure**, but I will expand a bit on this here.

The correlation matrix technique used by von Lucadou and a few others is based on the idea that mind–matter correlations might be interpreted as entanglement correlations (von Lucadou 2006, von Lucadou, Römer, & Walach 2007, Walach, von Lucadou, & Römer 2014). Along this idea, and given the postulate that entanglement correlations cannot be used for signal transmission, one consequence is that strict replication studies tend to fail, and no prediction can be made in what part of a system a mind–matter correlation may show up. According to von Lucadou and co-workers, one should give a system many degrees of freedom to increase the likelihood of such correlations appearing more often than would be expected by chance. The correlation matrix as has been used by von Lucadou et al. consists of two main ingredients:

(1) A number of physical and psychological variables are arranged in a table or matrix form, and for each intersection between a physical and psychological variable the corresponding correlation between these 2 variables is calculated (and can be entered in the corresponding matrix position).

(2) In the final analysis, the number of significant correlations in this matrix is counted and compared with the number of correlations that would have been expected just by chance.

With ingredient (2), it is possible to estimate a likelihood that the combined result (the number of observed correlations) has occurred by chance (under a null hypothesis).

It is worth pointing out that ingredient (1) can be seen as the process of defining and using a number of tests (correlations between variables in this case), whereas ingredient (2) resembles a sort of multiple analysis: A combined statistical measure is derived from a multitude of individual tests. As stated above, no prediction is made of which of the individual tests would be significant, but the combined statistics of all tests can finally be judged.

It is mainly this ingredient (2) that formed the basis of the analysis as defined for the study in this paper, in the sense that a number of different methods/tests are used. In particular, Analysis 1 is the classical analysis looking for mean shifts in the intended direction. Analysis 2 is different from this in that it looks at the distribution of the mean shift results from the 40 participants. It is possible that Analysis 2 would be highly significant (that is the distribution would deviate from the expected distribution under a null hypothesis), while at the same time the combined mean shift of all participants (which is precisely Analysis 1) would not be significant. Note that Analysis 1 is constituted by only a single statistical test, whereas Analysis 2 consists of 40 test results. No prediction is made on how the distribution of results in Analysis 2 may deviate from expectation. It was this principle that was inspired by the matrix method of von Lucadou et al.: To make no pre-

diction of precisely where a statistical deviation would occur, but rather to leave the system many degrees of freedom for deviations to show up. In the correlation matrix method, the number of significant correlations is counted and compared to the expectation value of a control dataset. In Analysis 2, a figure of merit is defined that describes numerically the deviation of the distribution from the expected distribution. This is just a more general form of how to combine the results of a multitude of tests. To make this point clear: This is not a replication of the correlation matrix method, in particular since the matrix elements in von Lucadou et al.'s experiments resemble correlations between psychological and physical variables (ingredient (1)). However, it resembles the idea of many degrees of freedom and applies it to a different kind of analysis (ingredient (2)).

One may argue that Analysis 1 would correspond to a classical analysis where a signal may be isolated from the data, whereas Analysis 2 would be more reasonable under the assumption of entanglement correlations as a putative explanation for significant effects. In this sense one may find it confusing to mix these two kinds of analysis. However, I would note two points with respect to this: First, Analysis 1 is still balanced between left and right intention, and the sequence of those intentions has been freely chosen by the participants. Therefore, if the chosen sequence would not be known, it would be impossible to derive a signal from the data (under the alternative hypothesis that the data has been influenced). In other words: Without knowing under which intention a stretch of data was generated, the computation of the result of Analysis 1 would not be possible. Second, one may view the combination of different types of analysis (like Analysis 1 and Analysis 2 here) just as an application of ingredient (2) of the matrix method as explained above: a case where multiple tests are done without predicting which one would be significant. It is also along this line that even more tests have been added, as for the cases of Analysis 3 and Analysis 4.

For Analysis 3 and Analysis 4, the number of stratistical tests (viewed as physical variables) was expanded from one to many. However, other than in the correlation matrix technique used by von Lucadou et al., the only correlation of the physical variables pertain to the left or right intention. This is a very reduced form of correlation and has only a loose connection to ingredient (1) of the correlation matrix method. However, the idea behind this was just to see if something surprising might happen, in that more correlations (or significant results) than expected might show up.

Pilot Study

No dedicated pilot study was conducted for this experiment, basically because the hardware and procedural design of this study seemed sufficiently

straightforward to make problems seem unlikely to occur throughout the data collection period. For the data analysis, the choice was to fully specify the analysis (as described in the sections above), and to put forward different kinds of analysis, though in a statistically sound way. Since one guiding idea was to make no prediction on which individual test or analysis would be significant, it was also not deemed necessary to conduct a pilot study with respect to the data analysis. Naturally, the study described in this paper might be regarded as a pilot study with respect to the design of new experiments but clearly not in the sense that in a typical pilot study the analysis might not be prespecified and subject to adjustment after the data had been looked at.

Exploratory vs. Confirmatory Analysis

Similar to the case for a pilot study, one may want to categorize the experiment at hand in terms of exploratory vs. confirmatory analysis. The label *exploratory analysis* seems often used to describe a process where a number of statistical tests is used on existing data, to find out which type of analysis might yield an interesting or unexpected result. A finding of interest may then be used as a hypothesis to test on new data, a process that then may be described as confirmatory analysis. Obviously it would be improper to report significant results of an exploratory analysis of this kind, without setting this into the context of all types of analysis that have been tried on the given database. Rightly so, this kind of practice might be the one most criticized.

The analysis done in this work is not exploratory in this sense, since the analysis has been prespecified before the data was looked at. It may be called exploratory only in the sense that a number of different analyses have been conducted, without predicting which one would yield a significant result. In this sense, the work is exploratory if the results would be used to generate new hypotheses to investigate in further studies. However, it should be pointed out that thinking along this line would imply that one has in mind to isolate one type of analysis, which then may show significant results on all future experiments of of this kind. According to von Lucadou, Römer, and Walach (2007) and Walach, von Lucadou, and Römer (2014), a confirmatory study that uses a single analysis that has been put forward from a former exploratory study may well fail. This is (according to those authors) due to the signal non-transmission theorem, and the decline effect that may be derived from it. It is in this sense that the study at hand has been designed to have many degrees of freedom in which correlations may show up.

Watt and Kennedy (2015) give a nice overview of exploratory vs. con-

firmatory analysis, and add the term *prespecified exploratory analysis*. Perhaps this might be an acceptable label for the study presented here.

Results Summary

To repeat the main results: Out of 4 predefined analyses, one was found to be significant with a probability $p = 0.0366$ that this result occurred by chance under a null hypothesis. The combined analysis of the 4 individual ones is found to be not significant, with $p = 0.2608$ to have occurred by chance under a null hypothesis.

A skeptical observer may say that the fact that the combined analysis is not significant means that no further discussion is necessary. If one's prior inclination is more to the end that psi may exist and may show up in experiments like this, then one may find the result of Analysis 2 at least interesting. The significant result for the distribution of the participants' individual results in Analysis 2 may yield a hypothesis for further study. To the knowledge of the author, no such investigation has been performed by other investigators. As stated above: Looking for the distribution of individual results rather than the significance of a combined result of many individuals has some similarity with the correlation studies where a statistical analysis is performed on the total number of significant correlations, without predicting which individual one would be significant. This direction of research may be supported by entanglement correlations in a generalized quantum theory (von Lucadou, Römer, & Walach 2007, Walach, von Lucadou, & Römer 2014, Atmanspacher, Roemer, & Walach 2002, and Filk & Römer 2011).

However, a followup experiment that would use only Analysis 2 might open up discussion about whether to view such an experiment as too strict a replication such that it might fail, or whether it may have sufficient internal degrees of freedom to allow for further significant results.

Acknowledgments

The author is grateful to M.A.P. for thoroughly reading the manuscript and for useful discussion. The author further thanks Walter von Lucadou and Eberhard Bauer for fruitful discussion and comments. Finally, I also thank the *Journal* reviewer for constructive criticism. This study was performed as private research by the author.

Notes

[1] Even if bias were not removed by the software RNG, this is a level of bias that would not be significant in the main experiment, because the main

experiment comprises 33 times fewer trials than this test run (288 million of the main experiment versus ~9,509 million of the test run). The bias detected with 9,509 million trials, in an experiment with 288 million trials would result in a z value of –0.46 and an insignificant cumulative chance probability of 32.5%.

[2] In three cases an error occurred on the hardware number generator during generation of control data of set a). This was caused by a minor bug in the program, which led to a low battery state and thus a low count rate on the number of zero crossings of the voltage comparator. The control data of the participants where this occurred was regenerated. It should be noted as well, however, that control data of set a) is not foreseen for analysis anyway.

[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 applications, as pointed out in a recent overview article on Bayesian and classical hypothesis testing (Kennedy 2014).

[4] Of course, in principle it would be possible to calculate the likelihood of the employed statistical tests analytically; however, a Monte Carlo approach was chosen here for simplicity and for better transparency of the data analysis. Furthermore, 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. A better approach in principle can be to use the existing dataset with random incursion points to generate the background distribution. However, in this case a problem might be the limited amount of available data.

[5] If the control set b) shows a significant deviation from randomness, it would be possible to subsequently generate more control datasets of type b) and/or more reference data to test whether the deviation would be systematic, or was a deviation by chance. If the deviation were systematic, the whole study would face an unforeseen problem, and probably no conclusions on the main experimental data could be drawn in this case.

[6] As can be seen from the statistical tests defined, some of the tests are performed on a combination of data under left and right intention, such that there would only be a single field for this column in a corresponding correlation matrix. However, this does not matter for the purpose here,

where all test data are combined by their individual probability rankings.

[7] Normally this would be a number describing the deviation of the test result from an assumed reference distribution. Since a ranking is applied subsequently, the reference distribution does not necessarily have to describe the exact expectation distribution of the test.

[8] The resulting 40 participant probabilities that each result occurred by chance in their corresponding class is an intermediary result here, which can be used to identify individual participants as deviating from the expectation value.

References Cited

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

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), 497–523.

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

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.

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *8*(1, Special Issue), 3–30.

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

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. *Psychological Bulletin*, *132*(4), 529–532.

Schmidt, H. (1987). The strange properties of psychokinesis. *Journal of Scientific Exploration*, *1*(2).

von Lucadou, W. (2006). Self-organization of temporal structures—A possible solution for the intervention problem. In Frontiers of Time. Retrocausation—Experiment and Theory, *AIP Conference Proceedings*, *863*, 293–315, Melville, New York.

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.

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.

Watt, C., & Kennedy, J. E. (2015). Exploratory vs. Confirmatory Study Analyses. Koestler Parapsychology Unit Study Registry, University of Edinburgh. http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/explore_confirm.pdf