REANALYSES OF GROUP TELEPATHY DATA WITH A FOCUS ON VARIABILITY

By Jan Dalkvist*, William Montgomery**, Henry Montgomery* and Joakim Westerlund*

ABSTRACT: Reanalyses of data from experiments on telepathic communication of emotions, as evoked by slide pictures, between groups of senders and groups of receivers are reported. In the present study, variability in performance rather than level of performance was in focus. Fits between variability in distributions of hits expected by chance and variability in empirical distributions were explored. The expected distributions were derived by means of the hypergeometric distribution, which provides the number of successes in a sequence of *n* draws from a finite population without replacement. Session level analyses showed that the variability in hit-rate was smaller than that expected by chance, particularly when the session groups who started as senders and those who started as receivers were analyzed separately and when the geomagnetic activity was low. Monte Carlo analyses indicated that these results could not be explained by stacking effects. Individual level analyses did not show any effects. In a second part of the study, the variability of responses to the individual target pictures was explored. The variability differed significantly among the pictures. Simulation showed that this effect was not attributable to stacking effects. Two predictions to be tested in an ongoing replication experiment are presented.

Keywords: telepathy, emotions, variability, hypergeometric distribution, simulation

The vast majority of ESP experiments have been performed and analyzed at the individual level. That is, data have been collected for each participant individually, and the unit of analysis has been the participant, even though the results in general have been summarized at the group level.

One reason why group experiments on ESP are relatively rare is probably the old and widespread opinion that group testing is inferior to individual testing in producing positive results (see, e.g., Rhine, 1947/1971). In line with this negative evaluation, several later studies have failed to produce any positive results (e.g., Haight, Weiner, & Morrison, 1978; Milton & Wiseman, 1999). Positive results have also been reported, however (Barker, Messer, & Drucker, 1975; Carpenter, 1988; Dalkvist & Westerlund, 1998), but attempts to replicate some of these results have failed (Carpenter, 1991; Westerlund & Dalkvist, 2004).

In any case, it would be premature to abandon group testing at this point in time. One reason is that, thus far, too few well-controlled group studies using different designs and types of ESP tasks have been tested to permit any definite assessment of the merits and drawbacks of group testing. For example, most studies have been concerned with clairvoyance or precognition and not with telepathy. Besides the above-mentioned studies by two of us (JD and JW), we know of only one group telepathy study (Auriol

et al., 2004). This long-term experimental series failed to demonstrate any deviation from chance expectation with respect to performance level, but performance variations among experiments that deviated significantly from chance expectation were found.

Another reason for continuing to use group testing is that this method is much less time-consuming than individual testing is. Thus, as long as we are not certain that group testing, in contrast to individual testing, will fail to uncover any ESP phenomena, group testing should be used for purely practical reasons. A further, less obvious, reason for not abandoning group testing in ESP research is that ESP may be critically dependent on social factors, such as the psychological atmosphere in a group of senders or receivers in a telepathy experiment.

Unfortunately, when running group experiments, one is faced with a big statistical problem, called "stacking," which probably has made many researchers refrain from doing group experiments. The problem is this: Due to the possible occurrence of dependency among participants' responses in group testing (e.g., due to the occurrence of a common response bias, such as a tendency on the part of the respondents to give one type of response at the beginning of a run and another type at the end of it), the statistical assumption of independent measures runs the risk of being violated. In general, the stacking effect acts to inflate the results by effectively reducing "*n*" in any (conventional) statistical test due to the occurrence of positive correlations among participants' responses caused by stacking (for example, when all participants invariably respond in exactly the same way, the effective *n* is reduced to one).

There are several ways of overcoming the stacking problem, however. One is by statistically correcting the data for the effects (Thoules & Brier, 1970), although this method is in general extremely laborious or uncertain, depending on the specific technique being used.

Another solution is to let the whole group of participants who have been subjected to the same experimental treatment be the measurement object in a statistical analysis, and not the individual participant, the rationale behind this method being, of course, that correlations among responses within groups become irrelevant by this procedure and that no stacking effect can occur among different groups because of the lack of communication among them. There are drawbacks to this method, however. One is practical. The method requires a considerable amount of data, that is, a large number of different groups. Another drawback is that the method is not generally applicable. It works well for analyses concerned with means or some other measure of the central tendency. However, it cannot be used reliably for analyses concerned with variability rather than the central tendency. This is because the response variation within groups, for statistical reasons, is reflected by the variation among corresponding group means (or central values of some other type). Thus, when considering, for example, the standard deviation of the means of a

performance measure for different groups of participants all of whom have been treated in the same way, we cannot tell to what extent it has been affected by within-session correlations among participants' responses that have been caused by stacking (in general, the variation will increase rather than decrease due to the occurrence of positive response correlations within the groups).

Still another possibility, which is free from any theoretical shortcoming, is to resort to a statistical simulation technique, a so-called Monte Carlo method (Dalkvist & Westerlund, 1998), where empirical data are compared to corresponding simulated data generated according to the null hypothesis using some appropriate random sampling technique and the set of empirical responses at hand. This method may, in effect, be useful as a complement to ordinary statistical methods, for example to check all significant results but omit all nonsignificant ones. By using such a selective strategy, the often time-consuming and technically demanding work required in doing simulations may be considerably reduced.

Since the spring of 1993, a series of group telepathy studies has been performed at the Department of Psychology, Stockholm University, with one of us (JD) as initiator. Based on the idea that strong emotional messages—for instance, signals of danger—may, for evolutionary reasons, be easier to transmit telepathically than are more neutral messages (Moss & Gingerelli, 1968), the studies have all been concerned with transmission of emotions as evoked by slide pictures.

As a first part of the present series of studies, five individual studies, which mainly served to generate a set of hypotheses (Dalkvist & Westerlund, 1998), were performed. These hypotheses were then tested in a comprehensive replication study (Westerlund & Dalkvist, 2004). The outcome of this study was clearly negative. However, a new finding, concerned with the order of sending and receiving telepathic messages, was obtained. To elucidate this finding, a reanalysis of previous data was carried out, leading to additional new hypotheses (Dalkvist & Westerlund, 2006). Still more new hypotheses were suggested by another reanalysis, and will be presented in the present paper. These hypotheses are concerned with variability in performance rather than with mean performance.

So far in the present project, only means have been used as a summary measure of performance. It should be borne in mind, however, that the mean (or any other measure of central tendency) describes only one particular aspect of the underlying distribution of measurements—its overall level. Another important aspect is the variability, as indicated, for example, by the standard deviation. Although a measure of the central tendency of a distribution of measurements and a corresponding measure of variability are not quite independent of each other—either mathematically or empirically (the standard deviation is, for example, most often positively related to the mean)—a measure of variability

may provide useful information over and above that provided by the central measure, as demonstrated, for example, by recent research on performance as related to aging (e.g., MacDonald, Nyberg, & Bäckman, 2006) and ADHD (Söderlund, Sikström, & Smart, 2007). Nevertheless, the specific information provided by the standard deviation (or some other measure of variability) is often neglected.

In some parapsychological contexts, variability in performance rather than the level of performance has been in focus. For example, in research on the decline effect, Carpenter found a decreasing run score variance, meaning that run scores started out either high or low at the beginning of the session but approached chance as the session progressed (e.g., Carpenter, 1966, 1968, 1969; Carpenter & Carpenter, 1967).

Another context in which the concept of variability has been considered is meta-analysis when tests are made to see whether different data sets in a large database are more heterogeneous (have greater variability) than expected by chance. If so, based on the assumption that deviating data sets tend to be less reliable than nondeviating ones due to systematic errors, deviating data sets are often discarded to make the database more homogeneous and therefore (it is assumed) more reliable (e.g., Honorton & Ferrari, 1989).

This procedure may be questioned, however. The argument is that greater heterogeneity than expected by chance may reflect real (i.e., parapsychological) effects rather than systematic errors, meaning that reducing the heterogeneity by discarding deviating data amounts to eliminating—or at least reducing—the very effects under study. There may, for example, be a bidirectional effect involved: While one part of the distribution may contain real hits, the opposite part may contain data resulting from psi-missing, that is, a reversed response pattern turning hits into misses in a systematic manner. For example, the finding of greater variation in hit-rate than expected by chance in ganzfeld data (e.g., Storm & Ertel, 2001) may be taken as evidence of a bidirectional effect, involving both hits and misses not expected by chance alone. Thus, rather than interpreting greater variability than expected by chance as a sign of errors, it can preferably be seen as suggestive of real effects, at least initially.

However, not only *greater* variability than expected by chance but also *lesser* variability than expected by chance may be taken as suggestive of a genuine effect. Such reduced variability may be expected to occur, for example, in very successful studies, where participants consistently perform at a high level. Conversely, reduced variability may also be expected to occur in studies where psi-missing occurs consistently. In either case, reduced variability is paired with a deviating central measure—a high one in a successful study and a low one when psi-missing predominates.

The main purpose of the present study was to reanalyze data from the above-mentioned group telepathy studies performed by two of us (JD and JW)—but now looking at variability.

Two different types of analysis were performed. One involved comparison between the empirical interindividual variability in hit-rate and the corresponding expected theoretical variability. The other type of analysis concerned the question of whether the stimulus targets differed from each other in interindividual response variability.

Before considering the present study, we will give a brief overview of the previous studies.

Previous Studies

The Original Studies

A total of 337 participants, 222 females and 115 males, with a mean age of 27 years, took part in the five original studies (Dalkvist & Westerlund, 1998). Most of the participants were undergraduate psychology students at Stockholm University, who chose to participate in the study as part of course requirements.

The studies comprised 24 single experiments in all, the number of experiments per study varying from two to nine. The mean number of participants per experiment was approximately 14.

As stimuli, 30 slide pictures were used, 15 with positive motifs (such as nature pictures and pictures of happy people) and 15 with negative ones (such as pictures of traffic accidents and starving children).

When the participants arrived at the laboratory, they were randomly divided into two groups, one sender group and one receiver group. The senders and the receivers were sequestered in two soundproof rooms, with one room in between. The two experimental rooms were connected by a signal device: a lamp in the receiver room that could be turned on and off from the sender room. There were two experimenters in the sender room and two in the receiver room.

The slides were presented in random orders, a new order for each group of senders. The senders' only task was to look at the pictures and to "hold on to" the feelings evoked by the respective pictures as long as they were being shown. The receivers were instructed to guess whether a given picture was positive or negative (they were informed about the number of slides, but not that the number of positive and negative pictures was the same). One of the experimenters in the receiver room watched the signal lamp and reported to the receivers when a new picture was being shown to the senders. Each picture was shown for 20 seconds, with an interstimulus interval of about half a second.

When all 30 pictures had been shown, the participants changed rooms, and those who had served as senders now served as receivers and vice versa.

Hit-rate, defined as number of correct responses or proportion of correct responses (when stimulus data were analyzed), was invariably used

as the dependent variable in the data analyses. Hit-rate was analyzed as a function of various personal and other factors.

The Replication Study

On the basis of the results of the five above studies, a number of predictions were formulated and tested in the replication study (Westerlund & Dalkvist, 2004). These predictions were all based on statistically significant (or, in one case, marginally significant) results obtained when data from the five studies were combined.

The new study was an exact replication of the latest of the five original studies, except that two additional minor control measures were adopted.

The replication study comprised 432 females and 173 males, 605 participants in all, with a mean age of 27 years. As before, the large majority of the participants were undergraduate psychology students at the Department of Psychology at Stockholm University, who chose to participate in the study as part of course requirements.

A total of eight predictions were tested. None of them was borne out (Westerlund & Dalkvist, 2004), which strongly argued against the possibility that some psi-phenomenon had been at work.

In spite of this failure, in a post hoc analysis, two physical moderator variables were entered: (a) local sidereal time (LST), an astronomical time and space measure, which is indirectly related to the magnitude of cosmic radiation that reaches the earth, and (b) disturbances in the global geomagnetic field, as measured by the *ap*-index. For a large number of different studies, performed in the northern hemisphere, Spottiswoode (1997) found both of these measures to be systematically related to the effect size of the studies. Our failures to replicate the previous positive results could not be explained in terms of differences in LST or *ap*-index, however.

A Follow-up Analysis

Although none of the eight predictions were born out, a significant unexpected result was obtained. In the original studies, a significant interaction effect was obtained between gender and receiver order, with an average hit-rate above expectation for the males when they started as receivers and an average hit-rate above expectation for the females when they started as senders. This interaction effect was not replicated in the follow-up study. Instead, a significant main effect of sender/receiver order was obtained, with a significant negative deviation from mean chance expectation for participants who started as receivers and a nearly significant positive deviation for those who started as senders.

Admittedly, this result was not predicted to occur, and as many as eight different predictions were tested in the study, meaning that the result did not reach significance when correction was made for number of tests. Nevertheless, inspired by earlier reports of effects of sender/receiver order in ganzfeld research (Haraldsson, 1980, 1985), we decided to carry out a follow-up analysis of the present sender/receiver order effect, based both on the original data set (with the first three studies, which were less well controlled, removed) and on data from the replication study (Dalkvist & Westerlund, 2006). In the following, we will collectively refer to data from the last two studies in the initial series of studies as the "old data" and data from the replication study as the "new data."

The analyses were not carried out at the individual level, as before, but at the group level. Specifically, each session (one of the two parts of a single experiment) was used as the unit of analysis, and session means were used as input in the statistical analyses, mainly to avoid the stacking effect, as discussed in the introduction.

Was the sender/receiver order effect obtained in the new study a real effect or was it attributable to sampling errors? In an attempt to answer that question, a systematic series of data analyses were carried out.

An initial finding was that the discrepancy between the old and the new data sets apparently could be explained in terms of geomagnetic fluctuations, which were much larger in the old study than in the new one. Moreover, the *ap*-index not only seemed to explain the difference in the sender/receiver order effect *between* the old and the new study, but also the sender/receiver order effect *within* each of the two data sets. Thus, independent of data set, the sender/receiver order effect turned out to be negatively related to the *ap*-index, although not significantly so in the case of the old data set.

As indicated by a step-wise multiple regression analysis on the whole data set, one additional variable was significantly related to the sender/receiver order effect, namely, a response style variable: number of negative guesses, which in contrast to the *ap*-index was positively related to the sender/receiver order effect. This relationship was weaker than that for the *ap*-index, however.

The Present Study

Interindividual Analyses

Data. Exactly the same data as those used in the above-mentioned study on the sender/receiver order effect (Dalkvist & Westerlund, 2006) were analyzed. Calculations were made for the old and the new study separately, as well as for both studies combined. The number of participants, sessions, and experiments over the old, the new, and the total data set are given in Table 1.

Table 1
Number of Participants, Sessions and Experiments
Over Three Data Sets

	Data		
Units	Old[a]	New	Total
Participants	240	605	845
Sessions	34	90	124[b]
Experiments	17	47	64

[a]The last two out of five studies.
[b]Four sessions were discarded because of technical failures.

In between-subjects analyses, the data were also divided into two subsets based on the activity of the geomagnetic field (GMF), one with high activity, as measured by the *ap*-index, and the other with low activity. The subset with high GMF activity included all experiments with a value above the median value plus a sample of 50% of the experiments falling exactly at that value. The subset with low GMF activity included all experiments with a value below the median value plus the remaining 50% of experiments falling exactly at that value.

Empirical Versus Theoretical Interindividual Variability in Hit-rate

General method. For each data set, analyses were carried out both at the individual level (for comparison purposes) and at the session level. To follow up the sender/receiver order effect, mentioned above, separate analyses were carried out for participants starting as senders and for participants starting as receivers.

Also, to eliminate errors associated with particular experiments, and thereby increasing the power of the analyses, in addition to analyzing session means, analyses were also conducted using the corresponding residuals around the means of the experiments (deviations of session means from the mean of any single experiment).

The general strategy was to compare empirical distributions of hit-rate with corresponding theoretical distributions, which assume that only random factors were at work, to see whether the two distributions differed from each other in variability. Comparisons between empirical and theoretical distributions were made using *F* tests. Because the expected distributions of hits would be obtained from an infinite number of respondents, the size of the data set expected by chance was assumed to be infinite; accordingly, an extremely high corresponding *df* value (10^6) was used in the *F* tests.

A theoretical distribution could not be constructed in a straightforward way, however, due to the procedure used in randomizing the stimuli, that is, sampling *without* replacement. If sampling *with* replacement had been used, the theoretical distribution would have been possible to obtain directly using the binominal theorem (although this method would have yielded less sensitive data, because the distribution of positive and negative pictures would then generally not have been optimal for discriminating between positive and negative stimuli, that is, containing the same number of positive and negative stimuli). However, as sampling without replacement was used, the empirical distribution became dependent on the number of positive and negative answers of each individual respondent. This problem was overcome by applying an appropriate algorithm, based on the so-called hypergeometric distribution, which can be used as a substitute for the binominal distribution when samples are drawn without replacement. However, because the participants' responses were assumed to be uncorrelated, possible stacking effects were not incorporated into the model. This problem was addressed by checking all positive results using simulations.

A computer program was written, in Java, to create the present type of distribution.

Theoretical distributions for individual data. The hypergeometric distribution is a discrete probability distribution that provides the number of successes in a sequence of n draws from a finite population without replacement. A typical example is the following: There is a shipment of N objects in which D are defective. The hypergeometric distribution gives the probability p that in a sample of n distinctive objects drawn from the shipment exactly k objects will be defective (Wikipedia, 2005).

The formula for the hypergeometric distribution may be written as follows:

$$p = \frac{D!(N-D)!n!(N-n)!}{N!k!(D-k)!(n-k)!(N-D-n+k)!} \qquad (1)$$

In the present case, the population is the 30 stimulus pictures (N), 15 of which are positive (D) and 15 negative (N-D). The parameter n is the number of responses in the less frequent response category (positive or negative responses). For example, if a participant has given 17 positive and 13 negative responses, n is equal to 13. The parameter k is the number of hits among the n responses in the less frequent response category. For example, if n = 13 and k = 7, there are 7 hits among a total of 13 minority responses. Insertion of these values into Equation 1 yields a p value of 0.27 of getting exactly 7 hits among 13 responses, all of which are negative (or positive if the positive responses are in minority).

Now, once n and k are known, the number of hits in the total set of N responses can be calculated. To show how by means of an example, let us assume again that $n = 13$ and $k = 7$. Let us further assume, as before, that the minority responses are negative. The 7 hits in response to negative stimuli then imply that the remaining 8 negative stimuli (15-7) are to be found among the 17 cases where the participant gave a positive response. The responses to these 8 negative stimuli will then be misses, whereas the remaining 9 positive responses will be hits. Thus, the total number of hits will be 7 + 9 = 16. More generally, the total number of hits (H) can be calculated from the following formula:

$$H = k + (N - n) - (D - k) \qquad (2)$$

Thus, by using the above two formulas, given the number of positive and negative responses, the expected probability for each possible number of hits can be calculated. Specifically, this is done by calculating p and H for each possible value of k for a given n. In the extreme case of $n = 0$, k is equal to 0, giving a hit-rate of $H = 15$ with a chance probability of 1, meaning that the only possible hit-rate is 15. The largest number of possible values of k and the largest number of different hit-rates is obtained when $n = 15$, giving a maximal hit-rate of $H = 30$, with a chance probability of $6.45 * 10^{9}$. In general, the possible number of different hit-rates, and hence the variability in hit-rate, decreases progressively as n decreases.

In constructing a probability distribution for a group of participants at the individual level, as a first step, a specific hit-rate distribution was constructed for each participant separately based on his or her number of negative/positive responses. Each such distribution gives a probability between 0 and 1 for each possible hit-rate to occur, with a total sum of 1. All individual distributions were then merged by summing the individual probabilities for each possible hit-rate. The distribution thus obtained was taken to be the expected probability distribution of hit-rates for the whole group.

Theoretical distributions for group data. In analyzing data at the group level, calculations were made to test whether the empirically obtained distribution of mean number of hits for groups of participants (for example, the groups that started as senders or the groups that started as receivers in each experiment) differed in variability from the corresponding distribution of mean hit-rates that would be expected if only random factors were at work.

In principle, such a theoretical probability distribution could have been obtained by combining every individual distribution in the group with all other distributions in the group to form an "average" distribution for the whole group. However, such a direct method would have required a great amount of computer time, because all possible hit-rates for one individual would have to be combined with all possible hit-rates for all other individuals.

To overcome this problem, a more effective method was developed. Theoretical probability distributions were thus constructed by using a procedure in which Equations 1 and 2 were applied recursively for a successively larger number of participants in a given group until a distribution of mean hit-rates for the whole group had been computed. As will be shown later, the procedure satisfies the necessary condition of giving the same results independent of the order in which data from the participants are entered into the calculations.

To calculate the expected probability distribution of mean number of hits for two arbitrarily chosen participants in the group, the probability distributions $p(H_1)$ and $p(H_2)$ across all numbers of hits H for participants 1 and 2, respectively, were calculated using Equations 1 and 2, as described above. A combined probability distribution for the two participants was then obtained by calculating the product $p(H_1)^* p(H_2)$ for all possible pairs of $p(H_1)$ and $p(H_2)$ for which both p values were greater than zero. Finally, a probability distribution across all different means of hit-rates H_1 and H_2 that could be formed was computed.

This computational procedure is illustrated in the following simple example involving only four hit possibilities, resulting from the low values of n. The hit probabilities greater than zero are assumed to be 0.24 for 13 hits, 0.52 for 15 hits, and 0.24 for 17 hits in Distribution 1 (would be true in the present case if $n = 2$), and .50 for 14 hits, and .50 for 16 hits in Distribution 2 (would be true in the present case if $n = 1$). The combined probability distribution for all pairs of hits in all six possible pairwise combinations of $p(H_1)$ and $p(H_2)$ will then be as follows:

.121 (13, 14)
.121 (13, 16)
.259 (15, 14)
.259 (15, 16)
.121 (17, 14)
.121 (17, 16)

and the corresponding probability distribution for the *means* of all 6 pairs of hits

.121 (13.5)
.121 (14.5)
.259 (14.5)
.259 (15.5)
.121 (15.5)
.121 (16.5)

(For example, the combined probability for the combination of 13 hits and 14 hits is 0.50 * 0.24 = 0.12, which is also the probability of the *mean* of 13 hits and 14 hits, that is, 13.5 hits.)

Finally, the probability distribution as a function of mean hit-rate is calculated by computing the sum of probabilities across all cases with the same mean hit-rate:

.121 (13.5)
.380 (14.5)
.380 (15.5)
.121 (16.5)

(For example, the probability for the mean hit-rate 14.5 is 0.121 + 0.259 = 0.380.)

To calculate the results for a subgroup (or whole group) that includes an additional third participant, the probability distribution of mean number of hits that already had been calculated for two of the participants was combined with the distribution calculated for the third participant using Equations 1 and 2, as before. When calculating the mean hit-rates for the new group of three participants, the distribution of mean hits for two participants was multiplied by two, because this distribution is based on twice as many participants as in the distribution for the single participant. Thus, mean hits were calculated according to the equation

$$M(H_1,H_2,H_3) = ((M(H_1, H_2)*2) + H_3)/3 \qquad (3)$$

where M is the arithmetic mean.

The probability for each $M(H_1,H_2,H_3)$ was calculated by combining the probabilities for the two distributions being used in the same way as when two single participants were combined. That is, we first calculated $p(H_1,H_2)*p(H_3)$ for all possible pairs of $p(H_1,H_2)$ and $p(H_3)$ for which p was greater than zero and then a probability distribution across all different means of hits that could be formed.

For groups with still higher numbers of participants, the procedure described above was applied recursively for each additional number of participants in the group. Generally, the mean hit-rates for a group of n participants was calculated according to the formula

$$M(H_1, \ldots, H_n) = ((M(H_1,\ldots,H_{n-1})*n\text{-}1) + H_n)/n \qquad (4)$$

and the probability associated with each $M(H_1,\ldots,H_n)$ by combining the probabilities for the distributions of H_1, $\ldots,H_{n-1}$ and H_n, respectively, as described above.

That the right-hand side of Equation 4 is indeed equal to $M(H_1, \ldots,H_n)$ follows from the fact that it can be reduced to $(H_1+\ldots+H_n)/n$, that is, the general expression for $M(H_1,\ldots,H_n)$. Obviously, this will be true independent of the order in which the participants are selected in the recursive procedure being used. Thus, Equation 4 will produce all possible $M(H_1,\ldots,H_n)$ for any order in which the participants are selected.

The final distribution was constructed in two steps: First, each theoretical distribution's probabilities were multiplied by the number of participants in the corresponding group of participants, to give a weight to each distribution in accordance with the number of participants. Second, all distributions were added together to form a total expected distribution.

As mentioned before, tests were also performed using residuals around the experiment mean instead of session means, to eliminate variation in hit-rate among experiments. Thus, an algorithm was also written to generate all possible results of residuals with respect to experiments (the differences of the results of the two session groups from the average results of the experiments) and calculating the probabilities of these results. The expected distribution of residuals obtained by chance was calculated in the following way: The probability of each pair of group results was calculated by multiplying the probabilities of the group results. The result of each pair was then added to the frequency of the residual result, which was equal to the average of the group results.

Analyses and results. The results of the comparisons between observed and expected variability for the old study are shown in Table 2. As can be seen from this table, there was no indication of there being any difference in variability between the empirical and the corresponding theoretical distributions. Thus, most of the *F* ratios obtained were close to one, and there was only one significant *F* ratio: that for the individual respondents starting as receivers, who showed a lower observed variation in hit-rate than was expected by chance.

Table 2
Comparisons Between Observed and Expected Variability, Old Data

Source of variation	*F*	Observed *SD*	Expected *SD*	*df*	*p*[a]
Individuals, total set	1.002	2.75	2.75	239	.960
Individuals, first receivers	1.320	2.39	2.75	122	.041
Individuals, first senders	1.260	3.10	2.76	116	.074
Means of sessions, total set	1.110	1.02	1.08	33	.634
Means of sessions, first receivers	1.490	0.88	1.08	16	.080
Means of sessions, first senders	1.110	1.17	1.11	16	.780
Residuals of sessions, total set	1.130	0.72	0.76	16	.824
Residuals of sessions, first receivers	1.140	0.72	0.77	16	.806
Residuals of sessions, first senders	1.140	0.72	0.77	16	.806

[a]Two-tailed

The results of the comparisons between observed and expected variability for the data from the new study are shown in Table 3. As can be seen from this table, there was a tendency for the empirical variability to be smaller than the theoretical variability—but only at the session level. At that level, two of the three *F* ratios obtained using ordinary means were significant: that for the total set of sessions and that for the sessions with participants starting as senders. The strongest results, however, were obtained for the residuals of the two sender/receiver order sessions. Taken separately, the results for the residuals of the sessions with participants starting as receivers and the sessions with participants starting as senders were highly significant. But the results for the residuals were also significant for all sessions taken together, although less strongly so. At the individual level, however, all of the three *F* ratios were close to zero and nonsignificant.

Table 3
Comparisons between Observed and Expected Variability, New Data

Source of variation	*F*	Observed *SD*	Expected *SD*	*df*	*p*[a]
Individuals, total set	1.05	2.68	2.75	604	.386
Individuals, first receivers	1.05	2.68	2.75	324	.575
Individuals, first senders	1.08	2.65	2.76	279	.368
Means of sessions, total set	1.43	0.95	1.13	89	.029
Means of sessions, first receivers	1.25	0.99	1.10	46	.344
Means of sessions, first senders	2.12	0.80	1.17	42	.003
Residuals of sessions, total set	1.76	0.60	0.80	42	.021
Residuals of sessions, first receivers	2.28	0.53	0.80	42	.001
Residuals of sessions, first senders	2.28	0.53	0.80	42	.001

[a]Two-tailed

The results of the comparisons between observed and expected variability over both data sets are shown in Table 4. The results shown in this table exhibit about the same pattern as the results for the data obtained in the new study. At the session level, however, the present *p* values are larger than those for the new study for all but one *F* ratio (between groups, first receiver), reflecting the high *p* values obtained at the session level in the old study.

Table 4
Comparisons Between Observed and Expected Variability, Both Data Sets

Source of variation	*F*	Observed *SD*	Expected *SD*	*df*	*p*[a]
Individuals, total set	1.03	2.70	2.75	843	.497
Individuals, first receivers	1.11	2.60	2.74	447	.132
Individuals, first senders	1.03	2.80	2.76	395	.674
Means of sessions, total set	1.30	0.97	1.11	123	.051
Means of sessions, first receivers	1.29	0.96	1.09	63	.191
Means of sessions, first senders	1.42	0.96	1.14	59	.084
Residuals of sessions, total set	1.54	0.63	0.79	59	.034
Residuals of sessions, first receivers	1.69	0.61	0.79	59	.011
Residuals of sessions, first senders	1.69	0.61	0.79	59	.011

[a]Two-tailed

In order to test whether the positive results obtained when comparing observed and theoretically expected variability at the session level were due to a stacking effect (or faulty calculations), a Monte Carlo method was used. Specifically, standard deviations were repeatedly recalculated for 100 simulated studies, obtained by substituting each original stimulus order in each study for a new randomized stimulus order. If stacking effects did occur they would become incorporated into the simulated data through the participants' responses (but not into the theoretical data, because the participants' responses were assumed not to be correlated in the present theoretical model). Hence, *p* values that were unaffected by stacking effects could be obtained by inserting empirical standard deviations into corresponding sampling distributions of simulated standard deviations.

Table 5 shows the *p* values obtained in this way for the whole data set and for the new one. As can be seen by comparing Table 5 with Tables 3 and 4, there was a very good agreement between the *p* values obtained by simulation and the corresponding *p* values obtained by comparing observed and theoretically expected variability. Thus, the possibility that the significant results obtained by comparing observed and theoretical variability at the session level were due to stacking effects could be refuted.

Table 5
P Values for Empirical Between-Session Standard Deviations of Mean Hit-Rates for the New and the Total Data Set Obtained by Monte Carlo Analyses Using 100 Simulated Studies

Source of variation	Data	
	Total	New
Means of sessions, total set	.03	.03
Means of sessions, first receivers	.05	.12
Means of sessions, first senders	.08	< .01
Residuals of sessions, total set	.01	< .01
Residuals of sessions, first receivers	.01	< .01
Residuals of sessions, first senders	.01	< .01

The results of the comparisons between observed and expected variability for the data associated with low activity of the earth's magnetic field are shown in Table 6. These results are very similar to those obtained in the new study. At the session level, the observed variation is thus smaller than expected by chance, particularly for the residuals of the two separate sessions, but not at the individual level.

Table 6
Comparisons Between Observed and Expected Variability, Low Geomagnetic Activity

Source of variation	*F*	Observed *SD*	Expected *SD*	*df*	*p*[a]
Individuals, total set	1.02	2.72	2.75	406	0.770
Individuals, first receivers	1.09	2.64	2.76	221	0.372
Individuals, first senders	1.02	2.77	2.75	184	0.908
Means of sessions, total set	1.50	0.91	1.12	61	0.048
Means of sessions, first receivers	1.42	0.94	1.12	32	0.215
Means of sessions, first senders	2.53	0.72	1.14	28	0.003
Residuals of sessions, total set	1.62	0.62	0.79	28	0.112
Residuals of sessions, first receivers	2.72	0.48	0.80	28	0.002
Residuals of sessions first senders	2.72	0.48	0.80	28	0.002

[a]Two-tailed

The results of the comparisons between observed and expected variability for the data associated with high activity of the earth's magnetic field are displayed in Table 7. There is a marked difference between these results and those for the data associated with low geomagnetic activity, just considered. There is still a tendency for the observed variation in hit-rate to be smaller than expected by chance, but this tendency is weaker than the tendency in the data for low geomagnetic activity, as indicated, for instance, by the lack of any significant *F* ratio.

Table 7
Comparisons Between Observed and Expected Variability, High Geomagnetic Activity

Source of variation	*F*	Observed *SD*	Expected *SD*	*df*	*p*[a]
Individuals, total set	1.05	2.69	2.75	436	0.511
Individuals, first receivers	1.14	2.56	2.74	224	0.173
Individuals, first senders	1.04	2.82	2.76	211	0.672
Means of sessions, total set	1.16	1.03	1.11	61	0.460
Means of sessions, first receivers	1.26	0.97	1.08	30	0.439
Means of sessions, first senders	1.08	1.11	1.15	30	0.822
Residuals of sessions, total set	1.48	0.65	0.79	30	0.183
Residuals of sessions, first receivers	1.47	0.65	0.79	30	0.185
Residuals of sessions, first senders	1.47	0.65	0.79	30	0.185

[a]Two-tailed

Using the same 100 simulated studies as above, Monte Carlo simulations were also performed to test the possibility that the difference between the low and the high geomagnetic activity results were due to stacking effects. As can be seen from Table 8, this possibility could be refuted.

Did the between-session standard deviations for the simulated studies differ from the corresponding theoretical standard deviation? To get an answer to that question, one-sample *t* tests were conducted as indicated in Table 9. As can be seen from this table, neither the means nor the residuals in any of the two data sets exhibited a significant difference. Thus, if there was any stacking effect, this effect was not large enough to affect the variability at the session level to a discernible extent.

Table 8

P Values for Empirical Between-Session Standard Deviations of Mean Hit-rates for Sessions With Low and High Geomagnetic Activity, Respectively, Obtained by Monte Carlo Analyses Using 100 Simulated Studies

Source of variation	Geometric Activity	
	Low	High
Means of groups, total set	.02	.17
Means of groups, first receivers	.10	.21
Means of groups, first senders	< .01	.34
Residuals of groups, total set	.04	.13
Residuals of groups, first receivers	<0.01	0.13
Residuals of groups, first senders	<0.01	0.13

Table 9

One Sample *T* Test of the Difference Between the Theoretical Between-Session Standard Deviation and the Mean of Between-Session Standard Deviations for 100 Simulated Studies

	New Data			
	SD (theoretical) - Mean *SD* (simulation)	*t*	*df*	*p*[a]
M	1.13 - 1.13 = 0.00	-0.05	99	.964
Residuals	0.80 - 0.79 = 0.01	-1.18	99	.239
	Both Data Sets			
	SD (theoretical) - Mean *SD* (simulation)	*t*	*df*	*p*[a]
M	1.11 - 1.12 = -0.01	0.98	99	.330
Residuals	0.79 - 0.78 = 0.01	-1.23	99	.222

[a]Two-tailed

Why did the session level analyses predominantly yield significantly smaller-than-expected variability in hit-rate while the individual level analyses did not? One possible explanation is that the smaller-than-expected between-session variation was compensated for by larger-than-expected within-session variation. Comparisons between theoretical and empirical within-session distributions contradicted this explanation, however.

Another possible explanation assumes that sessions differed with respect to their internal variability in hit-rate such that larger session groups had greater internal variability than did smaller session groups. In that case, the individual level variability would be larger than the session level

variability, because individuals in larger session groups would get lower weights than individuals in smaller session groups when the variation among session means was calculated. According to this explanation, there should be a positive correlation between a measure of the within-session variability in hit-rate and the size of the session group for the new and the total data set, but not for the old one, where no clear-cut difference between the individual level analyses and the session level analyses was obtained. As can be seen from Table 10, this prediction was borne out: Significant positive correlations were obtained between the standard deviation of the hit-rate scores within sessions and the number of participants in the session group.

TABLE 10
PEARSON CORRELATIONS BETWEEN NUMBER OF RECEIVERS AND WITHIN-SESSION STANDARD DEVIATIONS FOR THE OLD, NEW, AND TOTAL DATA SETS

Data set	*r*	*df*	*p*[a]
Old	-.10	32	.564
New	.33	88	.002
Total	.21	122	.021

[a]Two-tailed

To test whether the significant correlations between number of receivers and within-session standard deviation shown in Table 10 were genuine or attributable to some kind of stacking effect, the two correlations were repeatedly recalculated using, again, each of the 100 simulated data sets. For the new data set, 5 out of the 100 correlations were found to be larger than the empirical correlation ($r = .33$). That is, according to the present Monte Carlo simulations, the empirical correlation for the new data set was marginally significant. For the whole data set, however, as many as 12 of the simulated correlations turned out to be larger than the empirical correlation ($r = .21$), which thus did not reach significance. These findings indicate that an artifact, probably a stacking effect that was positively related to the number of participants in the session, did occur. Nevertheless, the almost significant simulated correlation for the new data set gives some support to the interpretation that sessions differed with respect to their internal variability in hit-rate such that larger session groups had greater internal variability than smaller ones.

Stimulus Target Analyses

General method. Thus far, we have focused on variability in overall performance, that is, general hit-rate, to see whether and how this variability differed from theoretical expectations under the assumption that only random factors were at work. In the present part of the study, we have

instead focused on variability in responses to the individual target pictures, to investigate whether and how the variability differed among them.

Interindividual response variability was measured at the session level. The receivers' responses were coded binarily. Specifically, a guess that a picture presented to the senders was positive was coded as "0" and a guess that the picture was negative as "1." For each target picture and session, the variability of responses was taken to be the standard deviation of type of guess (positive or negative). This measure takes on its highest value (= .50) when the numbers of positive and negative guesses are equal and its lowest value (= 0) when all guesses are either positive or negative.

According to the null hypothesis, the 30 target pictures do not differ with respect to variability. Using the above measure of response variation, this hypothesis can be tested statistically, using analysis of variance (ANOVA). (If the response variability had been measured on whole data sets instead of subsets, no such test could have been made, due to the lack of any error estimate.)

Analyses and results. To test whether the 30 stimulus pictures could be discriminated from each other in terms of the mean within-session standard deviation of positive and negative guesses, a one-way repeated measures ANOVA was performed with pictures as the independent variable and the standard deviation of the within-session guesses as the dependent variable for the old, the new, and the total data set. As can be seen from Table 11, a significant picture effect was obtained for the new data set and a nearly significant picture effect for the total data set, but no effect at all for the old one.

Table 11
Results From a One-way Repeated Measures ANOVA With Pictures as the Independent Variable and the Standard Deviation of Within-session Guesses as the Dependent Variables for the Old, New, and Total Data Sets

Study	*F*	*df*	*p*[a]
Old	1.01	29, 957	.449
New	1.59	29, 2581	.024
Total	1.37	29, 3567	.087

[a]Two-tailed

It is of some interest to note that tests corresponding to those above using ordinary session means instead of within-session standard deviations did not show any positive results at all.

Table 12 shows the results of simulating the above ANOVA analysis using one single simulated study. As can be seen, there was not even a tendency for the stimulus pictures to differ in any of the three data sets, thus

negating the possibility that the positive results shown in Table 11 were due to a stacking effect. (Given the absence of any effects at all in the simulated study, it was not worthwhile, we thought, to spend all the necessary time and effort to perform a full Monte Carlo analysis.)

TABLE 12
RESULTS FROM A SIMULATED ONE-WAY REPEATED MEASURES ANOVA WITH PICTURES AS THE INDEPENDENT VARIABLE AND THE STANDARD DEVIATION OF WITHIN-SESSION GUESSES AS THE DEPENDENT VARIABLES FOR THE OLD, NEW, AND TOTAL DATA SETS

Study	*F*	*df*	*p*[a]
Old	0.97	29, 957	0.512
New	0.63	29, 2581	0.937
Total	0.69	29, 3567	0.895

[a]Two-tailed

DISCUSSION

Empirical Versus Theoretical Interindividual Variability

In comparing empirical and theoretical interindividual variability in hit-rate, the strongest results were obtained when analyzing the residuals for the sessions with respect to the mean hit-rate of the corresponding experiment. Thus, using these residuals, for both the new and the total data set, the observed variability in performance was found to be significantly smaller than expected by chance in all three analyses, with a very low *p* value for the two separate sets of groups in the new data set. Similar results were obtained in the session level analyses for the new and the total data set using ordinary means, although some tests did not reach significance in that case. Monte Carlo simulations of the positive results indicated that these results could not be explained by the occurrence of stacking effects. Taken together, the findings suggest that the present analyses have revealed something interesting.

The present findings are related to previous results, showing a mean performance difference between groups of subjects starting as senders and groups of subjects starting as receivers (Dalkvist & Westerlund, 2006). The fact that the variability tended to be smaller than expected by chance in the present session level analyses sheds some light on this finding. Thus, the smaller-than-expected variability observed in the session level analyses indicates that the mean difference in hit-rate between the two sets of sessions is associated not with an increase in the variability among all groups, as might have been expected, but rather with *decreased* variability within each

of the two sets of sessions, reflecting the occurrence of coherence within each set of sessions.

The fact that residuals gave more strongly significant results than did the original session means reflects the fact that the mean hit-rates of the two sender/receiver order session groups were positively correlated across experiments in the new and the total data set. In the previous study just mentioned (Dalkvist & Westerlund, 2006), this correlation was utilized by using a paired samples *t* test instead of an ordinary *t* test to increase the power in comparing the two sender/receiver orders. In the present study, the correlation between the hit-rates of the two sender/receiver order session groups was instead utilized by eliminating the variation in performance among experiments using residual analysis. In principle, the two methods to decrease the error variance are analogous.

In contrast to the new and the total data set, the old data set did not yield any significant results at the session level. Again, this is in agreement with the preceding study (Dalkvist & Westerlund, 2006), which did not show any mean performance difference between the session groups starting as senders and the session groups starting as receivers in the old data set. As mentioned before, in that study, the difference between the old and the new data set could be related to variation in the geomagnetic activity (as was true in the present study as well), the old data being associated with greater geomagnetic variation than the new data. In accordance with that finding, only the data set associated with a low level of geomagnetic activity showed any significant results in the present study. This dependence of the results on the geomagnetic activity in both studies lends some credibility to the results.

In contrast to the session level analyses, analyses at the individual level did not generally show any significant results. This difference is puzzling, because the greater random variation at the individual level should, theoretically, be compensated for by a larger number of degrees of freedom and thereby yield equally powerful tests as those at the session level. This paradox could apparently be resolved, however, based on the fact that the within-session variability increased with the size of the session group in the new and the total data sets, even though this correlation apparently was partly due to a stacking effect. Given the positive correlation between the intrasession variability and group size for the new and the total data set, it follows logically that, for these two data sets, the between-session variability will be smaller than the within-session variability, because smaller weights will be assigned to individuals in larger groups than to individuals in smaller ones when session means are compared.

The above results obtained by comparing observed and theoretical variability are sufficiently strong to justify continued research. Specifically, in an ongoing replication study, we will test the following prediction:

> *Prediction 1.* When analyzed at the session level, the data will show smaller between-session variability in hit-rate

> than expected by chance for sessions with the same sender/receiver order, at least when the session means are replaced with the residuals calculated around the experiment mean.

This prediction will be tested using the same methods as those used in the present study.

Before turning to the target picture analyses, a comment should be made on the method used in the above analyses. Contrary to what many people believe, the methodological standards in parapsychology are in some respects higher than in other, comparable fields, for example, in using blind and double-blind protocols more often than is common in these fields (Sheldrake, 1999). Moreover, parapsychology has also contributed to the development of new methods. One example is the finding of a new statistical bias occurring when averages of responses affected by expectancies in some types of experiments are calculated (Dalkvist, Westerlund, & Bierman, 2002; Wackermann, 2002) and the suggestion of methods to avoid this bias (Dalkvist & Westerlund, 2006). The present method of constructing theoretical distributions based on the hypergeometrical distribution when sampling without replacement is used constitutes another example. When the method is used in group studies, however, one must make sure in some way that the results are not affected by stacking effects, as the model underlying the method assumes that the participants' responses are statistically independent.

In group studies, the present method of constructing theoretical distributions based on the hypergeometrical distribution should ideally be combined with a simulation method. Such a method guarantees that a statistically significant result is not caused by stacking, and can therefore, in contrast to the hypergeometrical method, also be used alone without any test of the stacking effect. However, simulations alone do not show whether or not a stacking effect does occur, and do not give any information about such an effect if it really does occur. But by comparing the two methods, it is possible both to establish and to characterize a stacking effect in terms of its strength and other properties of interest (for example, whether participants' responses tend to be positively or negatively correlated).

Of particular interest would be to investigate whether, and to what extent, the stacking effect is caused by response bias, as conventionally assumed, or reflects some genuine parapsychological effects arising within the group of receivers. This issue could perhaps be addressed by comparing receivers who are isolated from each other in time or space with receivers working in the same room at the same time using both simulation and the hypergeometrical method.

Interindividual Response Variability as Related to Target Pictures

Positive results were also obtained when response variability was related to target pictures using one-way repeated ANOVA, and the possibility that this finding was caused by a stacking effect was effectively ruled out by comparison with a simulated study. Accordingly, the following prediction was made for testing in the ongoing replication experiment:

> *Prediction 2.* A repeated measures ANOVA will show the 30 stimulus pictures to differ with respect to within-session variability in responses, as indicated by the mean standard deviation of the within-session type of response (positive or negative guess).

General Considerations

Taken together, from a strict empirical perspective, the results presented here are quite impressive. Had they been obtained in a mainstream study, one would surely expect at least some of them to be replicable. There is also some theoretical support for the present findings, however, namely from Carpenter´s (2004a, 2004b) recent first sight model, according to which psi phenomena emanate from deep unconscious processes. Most notably, consistent with this model, the differences in performance between participants starting as senders and participants starting as receivers, resulting in relatively low variability between session groups with the same sender/receiver order, might basically be an effect of priming, such that participants starting as senders were subliminally affected by seeing the pictures. This idea must be clarified and tested, however (it could, for example, be tested by relating senders' reported degree of emotional involvement in the pictures to their hit-rate). Nevertheless, considering the notorious difficulty of replicating positive results in parapsychology, we are far from certain that the present results will be replicable. However, even if the results turn out not to be replicable, we must still explain how and why the current findings were obtained.

As discussed in the introduction, exploring measures of variability may be very informative in suggesting the occurrence of specific underlying processes. However, before we know which of the above results, if any, are replicable, we will not attempt to interpret any of our findings in terms of such processes.

References

Auriol, B. M., Garcia, F., Puech, L., Lagrange, S., Morer, C., Campardon, M., et al. (2004). Agapé: Group telepathy. A long-term experimental series. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention,* 325–346.

Barker, P. L., Messer, E., & Drucker, S. A. (1975). Intentionally-deployed attention states: Relaxation. A group majority vote procedure with percipient optimization. *Proceedings of Presented Papers: The Parapsychological Association 16th Annual Convention*, 165–167.

Carpenter, J. C. (1966). Scoring effects within the run. *Journal of Parapsychology,* **30**, 73–83.

Carpenter, J. C. (1968). Two related studies on mood and precognition run-score variance. *Journal of Parapsychology,* **32**, 75–89.

Carpenter, J. C. (1969). Further study on a mood adjective check list and ESP run-score variance. *Journal of Parapsychology,* **33**, 48–56.

Carpenter, J. C., & Carpenter, J. C. (1967). Decline of variability of ESP scoring across a period of effort. *Journal of Parapsychology,* **31**, 179–191.

Carpenter, J. C. (1988). Quasi-therapeutic group process and ESP. *Journal of Parapsychology,* **52**, 279–304.

Carpenter, J. C. (1991). Prediction of forced-choice ESP performance: III. Three attempts to retrieve coded information using mood reports and a repeated-guessing technique. *Journal of Parapsychology,* **55**, 227–280.

Carpenter, J. C. (2004a). First Sight: Part one, A model of psi and the mind. *Journal of Parapsychology,* **68**, 217–254.

Carpenter, J. C. (2004b). First Sight: Part two, A model of psi and the mind. *Journal of Parapsychology,* **69**, 63–112.

Dalkvist, J., & Westerlund, J. (1998). Five experiments on telepathic communication of emotions. *Journal of Parapsychology,* **62**, 219–253.

Dalkvist, J., Westerlund, J., & Bierman, D. J. (2002). A computational expectation bias as revealed by simulations of presentiment experiments. *Proceedings of Presented Papers: The Parapsychological Association 45th Annual Convention,* 62–79.

Dalkvist, J., & Westerlund, J. (2006). Telepathic group communication of emotions: Announcement of predictions for an ongoing experiment. *Proceedings of Presented Papers: The Parapsychological Association 49th Annual Convention*, 314–319.

Haight, J., Weiner, D., & Morrison, M. (1978). Group testing for ESP: A novel approach to the combined use of individual and shared targets. *Proceedings of Presented Papers: The Parapsychological Association 24th Annual Convention*, 96–98.

Haraldsson, E. (1980). Confirmation of the percipient-order effect in a plethysmographic study of ESP. *Journal of Parapsychology,* **44**, 105–124.

Haraldsson, E. (1985). Perceptual defensiveness, ganzfeld and the percipient-order effect: Two experiments. *European Journal of Parapsychology,* **6**, 1–17.

Honorton, C., & Ferrari, D. C. (1989). Future telling: A meta-analysis of forced-choice precognition experiments, 1935-1987. *Journal of Parapsychology,* **53**, 281–308.

MacDonald, S. W. S., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences*, **29**, 474–480.

Moss, T., & Gengerelli, J. A. (1968). ESP effects generated by affective states. *Journal of Parapsychology,* **32**, 90–100.

Milton, J., & Wiseman, R. (1999). A meta-analysis of mass-media tests of extrasensory perception. *British Journal of Psychology,* **90**, 235–240.

Rhine, J. B. (1971). *The reach of the mind.* New York: William Sloane. (Original work published 1947)

Sheldrake, R. (1999). How widely is blind assessment used in scientific research? *Alternative Therapies,* **5**, 388–391.

Söderlund, G., Sikström, S., & Smart, A. (in press). Listen to the noise: Noise is beneficial for cognitive performance in ADHD. *Journal of Child Psychology and Psychiatry.*

Spottiswoode, J. (1997). Apparent association between effect size in free response anomalous cognition experiments and local sidereal time. *Journal of Scientific Exploration*, **11**, 109–122.

Storm, L., & Ertel, S. (2001). Does psi exist? Comment on Milton and Wiseman's (1999) meta-analysis of ganzfeld research. *Psychological Bulletin,* **127**, 424–433.

Thouless, R. H., & Brier, R. M. (1970). The stacking effect and methods of correcting for it. *Journal of Parapsychology,* **34**, 124–128.

Wackermann, J. (2002). On cumulative effects and averaging artefacts in randomised S-R experimental designs. *Proceedings of Presented Papers: The Parapsychological Association 45th Annual Convention,* 293–305.

Westerlund, J., & Dalkvist, J. (2004). A test of predictions from five studies on telepathic group communication of emotions. *Proceedings of Presented Papers: The Parapsychological Association 47th Annual Convention,* 269–277.

Wikipedia (2005). Hypergeometric distribution. Retrieved December 20, 2005, from http://en.wikipedia.org/wiki/Hypergeometric_distribution.

**Department of Psychology*
Stockholm University
106 91 Stockholm, Sweden
jd@psychology.su.se

***Department of Psychology*
Gothenburg University,Box 500
405 30 Gothenburg, Sweden

Acknowledgment

We would like to acknowledge the financial support of the John Björkhem Memorial Foundation.

Abstracts in Other Languages

French

RE-ANALYSES DES DONNEES DE TELEPATHIE
EN GROUPE AVEC UN FOCUS SUR LA VARIABILITE

RESUME : L'article présente des ré-analyses des données provenant d'expérimentations sur la communication télépathique des émotions, évoquées par des images sur des diapositives, entre des groupes d'émetteurs et des groupes de receveurs. Dans la présente étude, la variabilité dans la performance est plus centrale que le niveau de performance. Elle explore les accords entre la variabilité des distributions des succès tels qu'attendus par la chance seule et la variabilité des distributions empiriques. Les distributions attendues furent dérivées des moyennes de la distribution hypergéométrique, qui donne le nombre de succès dans une séquence de n tirages pour une population finie sans remplacement. Des analyses au niveau de la session montrent que la variabilité dans le taux de réussite était plus petite que celle attendue par la chance seule, en particulier lorsque les sessions des groupes débutant comme émetteurs et de ceux débutant comme receveurs étaient analysées séparément, et que l'activité géomagnétique était basse. Les analyses de type Monte Carlo indiquent que ces résultats ne peuvent pas être expliqués par des effets d'empilement. Les analyses au niveau individuel ne montrent aucun effet. Dans une seconde partie de l'étude, la variabilité des réponses aux images cibles individuelles est explorée. La variabilité diffère significativement entre les images. La simulation montre que cet effet n'était pas attribuable à des effets d'empilement. Deux prédictions à tester lors d'une expérimentation de réplication en cours sont présentées.

Spanish

REANÁLISIS DE DATOS DE TELEPATÍA GRUPAL CON UN FOCO EN LA VARIABILIDAD

RESUMEN: Reanálisis de los datos provenientes de experimentos relacionados con comunicación telepática de emociones, evocadas por

fotos en diapositivas, entre grupos de emisores y grupos de receptores fue reportado. En el presente estudio, la variabilidad en el desempeño mas que el nivel de desempeño, fue el foco de estudio. La concordancia entre variabilidad en distribuciones de aciertos esperados por azar, y la variabilidad en las distribuciones empíricas, fue explorada. Las distribuciones esperadas fueron derivadas por medio de la distribución hipergeométrica, que provee el numero de éxitos en una secuencia de n intentos a partir de una distribución finita sin reemplazo. Análisis del nivel de la sesión mostró que la variabilidad en el rango de aciertos, fue mas pequeña, que lo esperado por azar, particularmente cuando los grupos de sesiones que comenzaron como emisores y los que comenzaron como receptores, fueron analizados separadamente y cuando la actividad geomagnética era baja. El análisis de Monte Carlo indicó que estos resultados no podrian ser explicados por "stacking effect" {termino que hace referencia a puntajes espuriamente bajos o altos en un test de PES, debidos a una relación fortuita ocurrida entre los sesgos al adivinar de los percipientes y las peculiaridades de la secuencia de los objetivos (nota del traductor)}. Análisis a nivel individual no han mostrado presencia de este efecto. En una segunda parte del estudio, la variabilidad de las respuestas para las imágenes objetivo individuales fue explorada. La variabilidad defirió significativamente en estas fotos. La simulación mostrada en este efecto no fue atribuible a efecto staking. Dos predicciones que serán probadas en un experimento de replicación, en curso, son presentadas.

German

REANALYSEN VON DATEN BEI
GRUPPENTELEPATHIE FOKUSSIERT AUF VARIABILITÄT

ZUSAMMENFASSUNG: Es werden Reanalysen von Daten bei Experimenten zur telepathischen Übermittlung von Emotionen, evoziert durch Diabilder, zwischen Gruppen von Sendern und Gruppen von Empfängern berichtet. In der vorliegenden Studie stand eher die Variabilität der Trefferleistung als das erreichte Leistungsniveau im Mittelpunkt. Übereinstimmungen zwischen der Variabilität in den Trefferverteilungen unter Zufallsbedingungen und der Variabilität in den empirisch gefundenen Verteilungen wurden untersucht. Die erwarteten Verteilungen wurden mittels der hypergeometrischen Verteilung abgeleitet; mit deren Hilfe lässt sich die Anzahl der Treffer in einer Sequenz bei n-Ziehungen ohne Zurücklegen aus einer endlichen Population berechnen. Analysen des Trefferniveaus während der Sitzungen ergaben, dass die Variabilität der Trefferrate geringer ausfiel, als unter Zufall zu erwarten war, besonders wenn die Sitzungen der Gruppen, die als Sender begannen und diejenigen, die als Empfänger begannen, getrennt ausgewertet wurden und die geomagnetische Aktivität gering war. Monte-Carlo-Analysen ergaben, dass diese Resultate nicht durch Stacking-Effekte

erklärt werden konnten. Analysen einzelner Sitzungsverläufe ergaben keinerlei Effekte. Im zweiten Teil der Studie wurde die Variabilität der Reaktionen auf einzelne Zielbilder untersucht. Die Variabilität zwischen den Bildern unterschied sich signifikant. Eine Simulation ergab, dass sich dieser Effekt nicht auf Stacking-Effekte zurückführen ließ. Es werden zwei Vorhersagen gemacht, die in einem laufenden Wiederholungsexperiment überprüft werden.