

INVESTIGATING METHODOLOGICAL ISSUES IN EDA-DMILS: RESULTS FROM A PILOT STUDY

BY STEFAN SCHMIDT, RAINER SCHNEIDER, MARKUS BINDER,
DAVID BÜRKLE, AND HARALD WALACH¹

ABSTRACT: This article addresses methodological issues in current EDA-DMILS research. The authors conducted an exploratory DMILS pilot study with 26 sessions and found a medium size effect. No significant psi finding is claimed, as the experiment was not designed to find or investigate any; instead, the authors studied the variation of this effect as a function of different methodological approaches. They compared different EDA parameters (tonic or phasic) and found effects similar in size. Furthermore, they contrasted the traditional parapsychological way to parameterize data with methods adapted from psychophysiological research. These new parameters clearly outperformed the parameter used so far in DMILS/Remote Staring research. Respiration is correlated with the electrodermal system. The authors investigated whether EDA responses that are caused by sudden irregular patterns in breathing are also part of the remote intention effect. Therefore, they discarded all these responses from the data. The effect size dropped down to 30%-77% (depending on the parameter) of its original value. This indicates that the remote intention effect is very likely to also affect the pulmonary system. Finally, they compared different statistical methods for the evaluation of EDA-DMILS data and show the traditional percentage influence score method is neither appropriate nor suitable. The authors present alternative methods, and compare their power using the pilot study data.

Direct mental interaction between living systems (DMILS) is a standard experimental procedure for the investigation of the possibility that one system or person intentionally influences another system's or person's physiology from a distance. The majority of the DMILS studies include electrodermal activity (EDA) as the dependent variable; they are referred to as *EDA-DMILS* experiments. At least 30 EDA-DMILS experiments were conducted through 1997 (see Schlitz & Braud, 1997, for an overview). This figure also includes experiments on *Remote Staring*, a similar experimental paradigm that investigates whether a person's physiology can "detect" a

¹ This work is funded by the Institut für Grenzgebiete der Psychologie und Psychohygiene, Freiburg e.V. We are especially grateful to Florian Schaefer, who was particularly helpful to us in understanding and working with EDA. We also thank Werner Ehm, J. Schulte-Mönting, and Emil Boller, who helped us to understand the statistical procedures. We also are grateful to Robert Bishop and Fiona Steinkamp for comments on the article.

distant gaze. Schlitz and Braud's (1997) overview of the results of all EDA–DMILS studies and all Remote Staring studies shows that 14 out of 30 experiments have yielded significant results. The studies show nonhomogeneous effect sizes, with a mean of .25 (Rosenthal's r) for both research paradigms.

Because EDA is the only dependent variable in most of the studies described by Schlitz & Braud (1997), the recording, processing, parameterization, and evaluation of the EDA data is the crucial point in interpreting them. Therefore, Schmidt and Walach (2000) evaluated the EDA methodology of all published EDA–DMILS and Remote Staring studies and compared them with a sample of studies published in leading psychophysiological journals. The results of this evaluation indicated that the EDA methodology applied by parapsychologists did not compare to state-of-the-art EDA measurement recommended by psychophysiologicals (Boucsein, 1992; Venables & Christie, 1980). None of the studies complied with the "Publication Recommendations for Electrodermal Measurements" issued by leading psychophysiologicals in 1981 (Fowles et al., 1981), and most of them even violated common psychophysiological knowledge.

Therefore, we rearranged our EDA methodology in the DMILS facility in Freiburg, Germany, in collaboration with the Department of Psychological Physiology of the University of Wuppertal, Germany. When we modified procedures according to psychophysiological standards, new questions about the DMILS paradigm arose, namely, the following.

1. Which component should be investigated within the DMILS paradigm considering EDA itself is not a unitary phenomenon and consists of two components? For one of the components (the fast-changing phasic component) there are also different parameterization possibilities. Which parameter will be the most beneficial for DMILS research?
2. EDA as an indicator of autonomic activity is strongly related to the pulmonary system. Irregularities in respiration often result in so-called *electrodermal responses* (EDRs). Should these EDRs be regarded as artifacts and be discarded from the data? Or, on the other hand, is the reported DMILS effect due to irregularities in respiration that can be measured indirectly in EDA data?
3. There are different statistical procedures used in the DMILS/Remote Staring literature. Is there a difference among the techniques? What are the advantages and disadvantages of the different methods, and which technique is the most appropriate?

Because our knowledge regarding these questions was insufficient, we decided to look for empirical answers before conducting formal DMILS experiments, and therefore we performed two pilot studies.

PILOT STUDY

Our pilot study had a twofold aim. First, because our laboratory had been modified in many ways, one aim of the study was for the researchers working in the laboratory (Stefan Schmidt, Rainer Schneider, and Markus Binder) to get acquainted with the equipment and procedures and to identify potential problems. Second, we were looking for empirical answers to the questions formulated above.

Therefore, we designed the pilot study as an exploratory experiment. No specific hypothesis was formulated, and no significance testing was performed. Thus, it was not conceived as a publishable study to be included in any meta-analysis on DMILS studies. A detailed protocol was deposited before the start of the study. We conducted two pilot experiments in parallel and evaluated the data of all sessions ($N = 27$) together to increase power.

Participants

Participants were recruited by word of mouth among the members of our institutions and with friends. Some pairs were drawn from a sample of participants who had previously replied to a newspaper advertisement. Twenty-seven failure-free sessions were preplanned, as we expected problems with the new techniques. In total, 42 sessions with 84 participants were conducted. Fifteen sessions were discarded for various reasons, such as participants showing no spontaneous EDRs, failures in respiration recording, and so on. Finally, 27 sessions were deemed complete. One session, however, had to be discarded because of some recording problems that had not been detected earlier. Therefore, 26 sessions including 52 participants were entered into the analyses presented below. These participants had a mean age of 28.6 years ($SD = 7.1$), and 28 (54%) were male. Our participants were remunerated for their participation.

Apparatus

Skin conductance (SC) was recorded on the nondominant hand using the constant-voltage method (0.5V). Ag/AgCl electrodes 8 mm in diameter were filled with an isotonic electrode gel (TDE 246 Skin Conductance Electrode Paste, MED-Associates, Inc., St. Albans, Vermont, USA). This paste consists of 0.5 % NaCl in a neutral ointment base, according to the recommendations by Fowles et al. (1981). Electrodes were attached to the thenar and hypothenar eminences by means of double-sided adhesive collars at least 15 min before the start of data collection. The electrodes were connected to an isolated skin conductance coupler that treated the incoming signal by a time constant of 10 s. Thereby we obtained SC data in two channels reflecting the phasic component (skin conductance response [SCR]) and the untreated SC data.

Respiration was recorded by a strain gauge fixed with a belt at the participant's upper abdominal area. The belt was placed in a way that it recorded abdominal as well as thoracic respiration (Lorig & Schwartz, 1990). The strain gauge and the two EDA channels were connected to a measurement device (I-410 General Purpose System) that served as a multichannel bioamp and A/D converter.

Participants were placed in two different electromagnetically and acoustically shielded rooms located at 15 m distance. The agent's cabin contained a reclining chair and a computer monitor. The temperature was kept at roughly 22° C (72° F). In the receiver's cabin there was also a reclining chair and a computer monitor as well as a device (humidifier) to increase the cabin's humidity. Humidity was raised to a mean value of 50.5% at the start of the recording period. We tested different conditions for the most appropriate temperature. Because higher temperatures in the environment increase perspiration, nonspecific electrodermal responses are more likely to appear and tend to have higher amplitudes (Schaefer & Boucsein, 2000); therefore, an environmental temperature up to 26° C (79° F) is recommended (see Schmidt & Walach, 2000, p. 145). On the other hand, participants should feel comfortable during the experiment. We conducted a set of experiments with a temperature of approx. 23° C (73° F) and another set with a temperature of approximately 26° C (79° F). The overall average temperature was 24.9° C (76.8° F).

Procedure

We invited pairs of participants to our laboratory. We explained the ideas of our research and gave them a description of the subsequent experimental task. Then they had to decide who wanted to be the agent and who would be the receiver. The experimenter then fixed the electrodes at the receiver's hand to guarantee a minimum time lag of 15 min between attachment of electrodes and start of data collection (Schmidt & Walach, 2000, p. 145).

Thereafter, the experimenter escorted the participants to the two shielded cabins, explaining the equipment and procedures in detail and answering further questions. Then the receiver was connected to the recording equipment and seated in front of a computer monitor, which displayed a pleasant and colorful screen saver. Next, the agent was accompanied to his or her cabin, where he or she could see the receiver's SCR as feedback on a computer monitor. The experimenter locked all doors, went back to the lobby, which was situated between the two cabins, and started the experiment.

A session consisted of 20 epochs (10 activate epochs and 10 calm epochs), each lasting 1 min. During an activate epoch the German word for *activate* was inserted below the receiver's SCR curve in the agent's cabin. The agent's task was to activate the distant receiver by means of mentality or intentionality. For a calm epoch the instruction was in the opposite

direction. All influence epochs were interspersed by 15-s rest intervals. The sequence of activate and calm epochs was randomized for every single session by an algorithm drawn from a pseudorandom process in a special way to provide balanced sequence of the two conditions. The end of the experiment was indicated to the experimenter and the agent by a short sound generated by the computer. Participants then emerged from their cabins to share their experiences with the experimenter in the lobby.

Three of the authors served as experimenters in this study: Stefan Schmidt conducted 12 sessions, and Rainer Schneider conducted 5 sessions, and Markus Binder conducted 9 sessions.

Data Analysis

Experimental data were stored at a sample rate of 16 Hz in three different 12-bit channels (respiration, total SC, and SCR) with the experimental condition coded in a fourth channel. For further processing of the data we used software called EDR PARA Version 3.71 (Schaefer, 1999a). This software served several purposes. All data could be graphically displayed without seeing the experimental condition. For the calculation of skin conductance level (SCL), all 960 values ($60 \text{ s} \times 16 \text{ Hz}$) recorded in a 1-min epoch of the total SC channel were averaged and converted to the corresponding value of SC expressed in microSiemens (μS). This mean was recorded in an output file.

In the SCR channel the software automatically detected single SC responses and parameterized them for latency on the start of the recording epoch, rise time, amplitude, and recovery time (see Schmidt & Walach, 2000, p. 142, for details). All data were written to an output file. Each response could be observed graphically together with the covarying respiration activity and could interactively be discarded from the output file. All SCR data were filtered by a 0.5-Hz low-pass filter.

The software parameterized SCRs only when onset and peak were placed within the epoch. In so doing, SCRs starting at the very end of an epoch and peaking in the following rest period could not be detected. Therefore, we extended every experimental epoch by 2 s, thereby shortening the subsequent rest periods by 2 s. In so doing, SCRs with an onset at the end of an epoch peaking within 2 s after the end of this epoch could also be included in the analysis. Yet SCRs having both onset and peak within these 2 s were excluded automatically by a filtering mechanism. This procedure enforced the idea that the time of the onset, and not the time of the peak of any SCR, should be related to the remote intention.

A special software module called EDR Select Version 2.1 (Schaefer, 1999b) assigned the epochs of all sessions to the experimental conditions activate, calm, and rest. The output file of this software was analyzed with the statistical package SPSS for Windows 9.0.

Which Parameters Should Be Taken? Assessing Different EDA Components

EDA consists of two components. The fast-changing *phasic* component (SCR in this study) reflects responses to certain stimuli with a typical pattern of the EDA curve. It rises after a certain latency time to a certain peak and slowly recovers. The other component is a slow-changing *tonic* component (SCL in this study) that reflects the overall arousal of the participant. Because the DMILS setup does not include any overt stimuli whatsoever, only tonic parameters reflecting overall arousal were of interest but, apart from SCL, another tonic parameter can also be derived from the phasic component. This is because the typical SCR patterns can also be found in the absence of any stimulus. These are called *nonspecific responses* (NS.SCR). The number and size of these NS.SCRs also reflect overall arousal.

Both SCL and SCR have been dependent variables in DMILS experiments. Whereas the early experiments conducted by Braud and Schlitz (see Braud & Schlitz, 1989, for an overview) have emphasized the phasic component, the majority of researchers in the 1990s have assessed the tonic component (e.g., Delanoy, Morris, Brady, & Roe, 1999; Wiseman & Schlitz, 1999).

Obviously, any alleged DMILS/remote staring effect in the EDA could be only in the phasic component, only in the tonic component, or in both. From a theoretical perspective, the effect should be found in both. The question of which component would be more appropriate—or, in statistical terms, which would be more powerful for detecting this effect—remains unanswered. We therefore recorded both components simultaneously to compare their outcomes.

To do that, it was important to decide how the components should be parameterized. For the SCL the situation is simple, because most psychophysicists average SCL data over a certain time period (see Schmidt & Walach, 2000, p. 147, for details), and this was done in all DMILS/remote staring studies using the tonic component. However, with regard to the phasic component the situation is slightly different. This is depicted in Figure 1, which shows the raw data of a very high activity SCR recording of 1-min duration. The aim of the parameterization process is to find the most appropriate measure to describe the activity pattern that can be easily seen in the graph. There are at least three possibilities.

- *Number of NS.SCRs.* One could just count the number of responses exceeding a certain threshold within a certain time period. This is shown by the numbers 1–4 for the first four SCRs.
- *Sum of amplitudes of NS.SCRs.* Instead of just counting the unweighted SCRs, one could also take their size into account by summing the amplitudes of all SCRs exceeding a certain threshold within a certain time period. This is shown in Figure 1 by the dotted line indicating an amplitude of 1.55 μ S for SCR No. 3.

- *Mean over all recorded data points.* Whereas the first two possibilities are usually applied in psychophysiology, parapsychologists preferred a different solution. Almost all DMILS/remote staring experiments in which the phasic EDA component was used calculated a mean of all recorded data points within a certain time period. This is shown in Figure 1 by the shaded area indicating the mean value of 0.61 μ S.

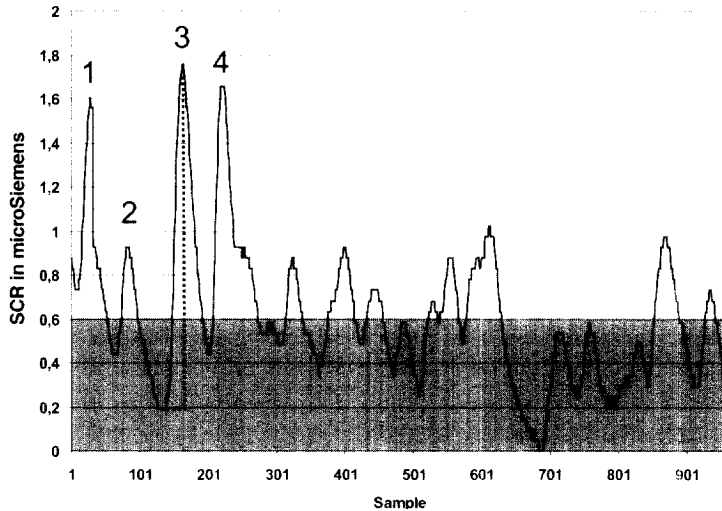


Figure 1. Nine hundred sixty samples (1 min) of skin conductance response (SCR) recorded from a participant.

Which one of these parameters provides the best statistical power to detect a purported DMILS effect is dependent on the way this effect operates. If the effect is responsible for the generation or absence of SCRs, then Method 1 or 2 may be the most appropriate, because they only take information from the SCRs, leaving the flat areas untouched. If, on the other hand, the effect acts in a general way, then the methods of averaging all data might have a higher power.

Also, for the comparison of number of NS.SCR and sum of amplitudes of NS.SCR different hypotheses regarding a purported DMILS effect can be formulated. If the DMILS effect is responsible for only the small SCRs, whereas the bigger ones are due to some other internal (e.g., cognitive) stimuli, then the unweighted number of NS.SCR should outperform the sum of amplitudes. If, on the other hand, the supposed effect is partly responsible for most of the SCRs, then the sum of amplitudes will have the better power to find the effect. Furthermore, for number of NS.SCR a threshold for SCRs has to be defined. Whereas a large

threshold will favor substantial responses, a small one will give a higher impact to very small fluctuations within the SCR data.

To find out which of those explanations are supported by our data, we applied each parameterization method to each of the 20 epochs of each session's data. Then we summed the values of all 10 activate and all 10 calm epochs for each session. In this way we obtained 26 pairs of activate–calm values for each parameter. Then we calculated a Wilcoxon signed rank test on these 26 pairs, obtaining z scores. We chose this analysis method beforehand because it was the most appropriate procedure. For further discussion and comparison of analysis methods, see the *How to Calculate?* section.

Results

Figure 2 shows the obtained z scores. All z values are positive, indicating effects in the intended direction (higher EDA activity for the activate condition). There seems to be almost no difference between the tonic and the phasic component parameterized according to psychophysiological standards. The corresponding effect sizes obtained by $r = z / \sqrt{N}$ is .42 for the tonic component and .41–.35 for the three phasic parameters number of NS.SCR (with the two different thresholds 0.015 μ S and 0.02 μ S) and sum of amplitudes.

The only exception is the parameter calculated as a mean of all recorded SCR data. The z score drops to 1.16 ($ES(r) = .22$), indicating that this traditional DMILS method has approximately only half of the power of other measures.

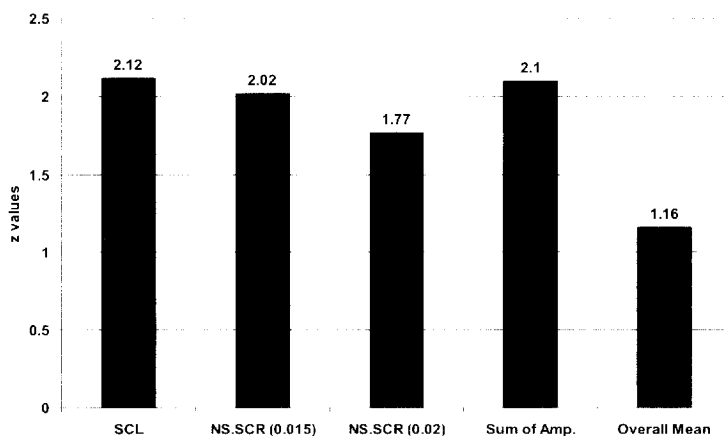


Figure 2. Results of the comparison of activate and calm epochs in electrodermal activity data. Differences are expressed as z scores and shown separately for the five different parameters that have been applied to the data. SCL = skin conductance level; NS.SCR = number of nonspecific skin conductance responses; Amp. = amplitude.

From the comparison of the three psychophysiological measures one can conclude that both large and small SCRs are affected. The fact that the sum of amplitudes shows a high z score can be interpreted as a strong influence of large SCRs, because they receive a higher weight in that measure, but the fact that the number of NS.SCR with the lower threshold show a slightly higher z score would indicate that especially the very small SCRs appear more often during activate than during calm epochs.

IS RESPIRATION AN ARTIFACT? ASSESSING THE INFLUENCE OF RESPIRATION ON EDA-DMILS DATA

The electrodermal system is strongly linked with the respiratory system (Boucsein, 1995). Figure 3 shows a typical pattern that can be found when the respiratory and the SCR recordings are simultaneously displayed.

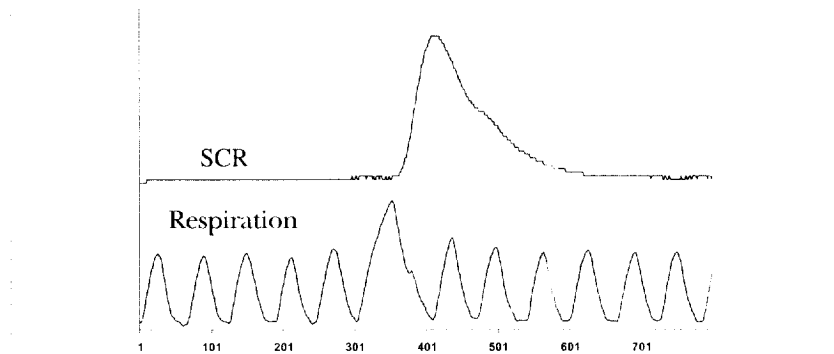


Figure 3. Eight hundred samples (50 s) of skin conductance response (SCR) and respiration data recorded from a participant.

Shortly after sample 300 the participant leaves her characteristic respiration rhythm and continues to inhale instead of the expected exhalation. This is indicated by the extended rising period in the respiration chart. After a latency time of approximately 2.1 s there is an onset of an SCR with an amplitude of 0.188 μ S. One can expect the onset of an SCR after a latency interval of 1–5 s after the start of irregular breathing. This response of the electrodermal system to irregular breathing is similar to the orienting response, in which an SCR is expected after a latency period of 1–5 s.

Knowing this, we can formulate interesting hypotheses regarding the manner in which a purported DMILS effect operates. The most significant findings in DMILS and remote staring studies have resulted from the use of EDA as a dependent variable. Studies that have used other

physiological measures, such as heart rate, blood pressure, blood volume pulse, skin temperature, muscle tension, or respiration frequency (Braud & Schlitz, 1991; Braud, Schlitz, Collins, & Klitch, 1985; Radin, Machado, & Zangari, 1998; Rebman, Radin, Hapke, & Gaughan, 1996; Wezelman, Radin, Rebman, & Stevens 1996), have shown only inconsistent, and few significant, results. These findings could be interpreted in two ways. First, the effect is global, but the changes are so small that it can only be found within the EDA, because skin conductance is for some reason the most appropriate variable. Second, the effect is local and affects only EDA, while other physiological systems, even with large samples and the best possible methodology, would never show this effect. If we cut these ideas down to the accessible EDA–respiration connection we can formulate the following hypotheses.

1. The purported DMILS effect is local and can be found only in the electrodermal system.
2. The purported DMILS effect is local in the respiration system. Because of irregular respiration eliciting SCR it can be found in the electrodermal system also.
3. The effect is global and affects at least two systems (respiration and EDA) simultaneously.

These hypotheses can be tested if we edit the phasic EDA data for respiration ‘artifacts’ (see, e.g., Figure 3). The SCR displayed in the graph is elicited by irregular breathing. If we discard all related SCRs from our data we get a second data set, from which all effects of irregular respiratory activity are eliminated. The comparison of the edited and unedited data provides insight regarding the three different hypothesis. If Hypothesis 1 were confirmed, the editing procedure would have only removed artifacts. Therefore, the effect in the edited data set should have been the same or could even slightly increase as the signal-to-noise ratio was improved by the removal of artifacts. If the second hypotheses were true, the effect should have dropped to zero, because the relevant information was deleted out. If Hypothesis 3 were true, the effect should remain the same or should slightly drop, because some part of the relevant information was removed. In the way the study was designed, and using our procedure, we could not expect an exact decision among these hypotheses, but by comparing edited and unedited data sets we achieved better understanding of the effect which, in turn, allows for more appropriate process-oriented research in the future.

The Procedure of Editing

Editing processes, like the one that was necessary for our data, can hardly be done automatically. A semiautomatic solution might be possible but was not available to us.

Therefore, we edited the data manually. This is rather troublesome, because manual editing processes are usually highly subjective and intuitive.

Unfortunately, the relation between EDA and respiration is not as unambiguous and linear as it is displayed in the example in Figure 3. For example, irregular breathing sometimes elicits SCRs, but sometimes it does not. On the other hand, strong SCRs may result in irregular breath-

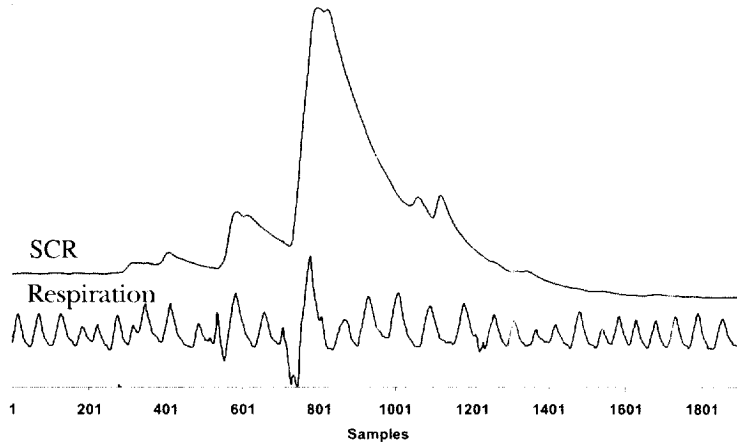


Figure 4. One thousand nine hundred ten (approximately 2 min) samples of skin conductance response (SCR) and respiration data recorded from 1 participant. The pattern of irregular breathing with simultaneous strong SCR shows the mutual interaction of the electrodermal and pulmonary systems.

ing, but they do not have to. Often one finds chains of mutual responses—see, for example, Figure 4, which presents 1,910 samples (approximately 2 min) of data in which several SCRs and several irregularities in breathing are interacting with each other. In searching for assistance in the literature we found that most psychophysicologists edit their SCR data for respiration artifacts themselves in a way that is not explicitly described (single-expert procedure). Confronted with these problems, we decided to develop an exact catalogue of criteria for data editing. The data of the pilot study were not used to develop this catalogue. We used data collected for a different study and data recorded for the testing of our equipment. The catalogue can be obtained from Stefan Schmidt. This catalogue had a twofold aim. Starting from our intuitive and subjective decisions the catalogue should formulate explicit rules for editing, guaranteeing a replicable process of decision making. Furthermore, the catalogue should provide clear solutions to difficult situations, such as those described earlier.

This work was done by Stefan Schmidt, Rainer Schneider, and Markus Binder. They started by discussing observed phenomena and their relations in the data. Then they identified problems and discussed different solutions. The solutions were checked with different data, and the most appropriate solution was entered into the catalogue. Next, each of them edited some data sets using this catalogue and then compared the results. By discussing the reasons for different decisions, they explicitly formulated any implicit decision criteria. They further discussed the advantages and disadvantages of different decision strategies and extended the catalogue accordingly. This procedure was repeated many times, until a sufficient convergence of the different ratings was achieved. The criteria and solutions for difficult situations were checked with Dr. Florian Schaefer, an EDA specialist.

Results

Stefan Schmidt, Rainer Schneider, and Markus Binder edited all 26 data sets. They were blind to the experimental condition during the editing process. The reliability over all three raters was .92 for the number of NS.SCR (reliability on all decisions combined by the three raters according to Bortz & Döring, 1995, p. 252) and .91 for the sum of amplitudes, respectively. For statistical analyses, they compared their ratings and discarded every SCR that was edited by at least two of the three raters to form a final edited data set. The results are shown in Figure 5. For all three phasic parameters z scores were lower when all respiration-linked SCRs were discarded from the data.

These results clearly show that SCRs elicited by respiration have a part in the total effect. Therefore, it is very likely that this effect cannot be reduced to only a local EDA phenomenon. The data presented favor Hypothesis 3 for a general effect on both physiological systems, followed by Hypothesis 2. However, the reported study was exploratory in nature, and therefore these findings should be confirmed by future studies.

HOW TO CALCULATE? ASSESSING DIFFERENT STATISTICAL PROCEDURES FOR DMILS DATA

The question of how to calculate significances or effect sizes from DMILS data is not a trivial one. Several different procedures have been applied in different studies, and we describe them shortly, presenting their advantages and disadvantages. We outline two methods that have not been applied before. We applied all procedures to our data and compared their outcomes.

Data sets of DMILS/remote staring experiments consist of several sessions with different participants. Each single session consists of

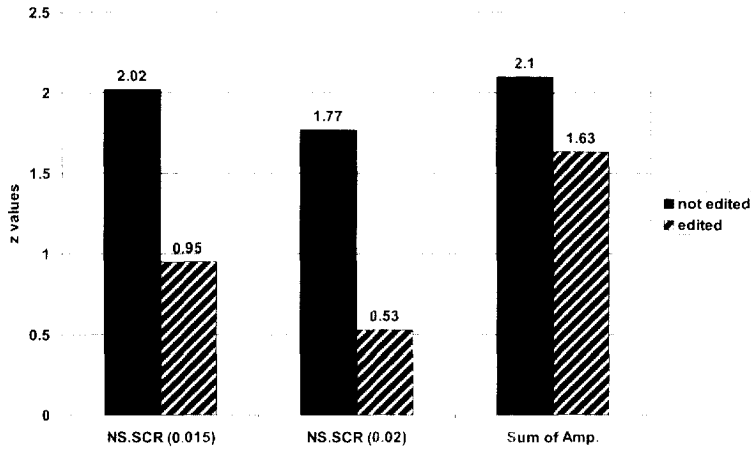


Figure 5. Results of the comparison of edited and unedited skin conductance response (SCR) data for the three parameters sum of amplitudes (Sum of Amp.) and numbers of nonspecific SCRs (NS.SCR), with minimum thresholds of either 0.015 μ S or 0.02 μ S. The unedited values show the result of the comparison of activate and calm epochs in the phasic electrodermal activity data, expressed in z scores. The edited values show the same results after the removal of all SCRs that were caused by irregularities in respiration.

several epochs reflecting the different experimental conditions (e.g., calm-activate, calm-control, stare-control) with the same number of epochs for both conditions. Within these epochs physiological data are recorded from a set of many data points. These data recorded within an epoch are combined to form a single parameter representing the amount of activity in each epoch. Different possibilities for computing this parameter were described in the *Which Parameters Should be Taken? Assessing Different EDA Components* section. The statistical evaluation must be aware of this triple dependency on the data (a) on the level of measurement points, (b) on the level of epochs, and (c) on the level of participants: (a) and (b) are dealt with by aggregation (i.e., summing or averaging), and (c) has to be taken into account by the statistical procedure used for testing.

PERCENTAGE INFLUENCE SCORE

The percentage influence score (PIS) was developed by Braud and Schlitz (1991, p. 5). It is calculated by dividing the sum of all epochs of one experimental condition by the sum of all epochs for both conditions for each session. One thus obtains a PIS for each session. In the absence of any effect the PIS is 50%, and each deviation from that mean reflects

an experimental effect. For significance testing all sessions' PIS are tested with a single mean *t* test against the expected mean for the null hypothesis (i.e., .50).

Advantage

The PIS is standardized by the mean. An indicator of success can be easily calculated for each session and can be easily communicated and explained to participants. The procedure is applicable to all kinds of dependent variables.

Disadvantage

The PIS is not standardized by the standard deviation. Therefore, PIS scores do not reflect experimental success. A large effect size (with a very small standard deviation in the original data) can result in a PIS of 50.01%, whereas a very small effect size (with a large standard deviation) can lead to a PIS well over 60%. Interpretation or comparison of different PISs says nothing about effects in the data. Furthermore, the statistical power of the PIS is highly dependent on the ratio of the standard deviation and the initial value.

This is demonstrated by a set of randomly generated data simulating DMILS data. We randomly drew a set of 30 numbers from a normal distribution (Variable 1). With the help of a second set of random numbers we calculated another set of 30 numbers (Variable 2) that is highly correlated ($r = .9$) with Variable 1 (see Boller & Schweizer, 1992, for details). To simulate a DMILS effect we added a constant to Variable 2. Variables 1 and 2 are now very similar to a DMILS data set consisting of 30 sessions, with Variable 1 representing the sum of all calm epochs and Variable 2 representing the sum of all activate or control epochs, respectively. For the calculation of PISs we systematically added a constant value (either 2, 3, 5, 10, 20, 50, or 100) to every single data point in the set and received seven new data sets with the basic level shifted by the respective constant. We then calculated a paired *t* test, a PIS, and a single mean *t* test for the PIS of each of these data sets. The results are shown in Table 1.

TABLE I
 INFLUENCE OF A CONSTANT VALUE ADDED TO A
 VIRTUAL DATA SET ON TWO DIFFERENT TEST STATISTICS
 (PAIRED *t* TEST AND SINGLE MEAN *t* TEST)

	2	3	5	10	20	50	100
Paired <i>t</i> test	2.871	2.871	2.871	2.871	2.871	2.871	2.871
<i>t</i> score of single mean <i>t</i> test	1.838	2.191	2.497	2.697	2.787	2.839	2.855
PIS	53.25	51.75	51.04	50.54	50.27	50.11	50.06
<i>p</i> value for PIS	.076	.037	.018	.012	.009	.008	.008

Note. Percentage influence scores (PISs) and *p* values (two-tailed) of the single mean *t* test scores reflect the change produced by the added value.

As one can see in the table, the statistical power to detect the effect of the PIS procedure is highly dependent on the constant that is added to the variables. The basic structure of the data is always the same, as reflected by the constant *t* value for the paired *t* test over all conditions. However, with a higher constant added the ratio of the standard deviation (always the same) and mean (shifted by the constant) changes. The PIS score is dropping toward the 50% value, and the *t* value of the corresponding single mean *t* test slowly approaches the expected value of 2.871. Likewise, the higher the constant added, the more the power of the *t* test improves.

In other words, the chances for a DMILS researcher to detect an existing effect in such data is highly dependent on the luck he or she has when he or she calculates the summed values for the epochs. Those who take, for example, the mean and convert it to the unit of skin conductance will end up with low values (e.g., 1.79 μ S per epoch). They might find nice PIS values in the 55% area or even higher, but the corresponding *t* test has only low power to detect the effect. This situation is made worse by high variances in the outcome values (see *Results* section), but those who just sum the bit values from exactly the same data will obtain high values (e.g., 2,400,000 per epoch) with correspondingly low PIS scores (in the 50.001 range) and a powerful test of significance. Unfortunately, this unreliable statistic has been used in almost all DMILS and remote staring studies.

TWO-COMPONENT MODEL

The model we now describe was named the *two-component model* by us because it combines two statistical tests. In the first stage, after the epochs are parameterized, a Wilcoxon signed rank test is calculated for every session, with pairs of epochs of two different consecutive conditions serving as the paired data for the test (e.g., for a session with 20 epochs in

sequence of ABBABBAABAABBAABAABB, A-B, B-A, B-A, B-A, B-A, A-B, B-A, A-B, A-B, A-B are forming 10 data pairs entering the Wilcoxon test). Thus, a z value can be assigned to every session. In a second stage, a test of significance of the whole experiment is applied. The z scores are either combined by Stouffer's Z method or are tested by applying a single mean t test against 0.

The method of calculating a Wilcoxon test on all epochs of one session gives a different priority to outliers other than just summing the epochs over the whole session. Imagine that one of the epochs contains data 10 times higher than in the other epochs. If the data are just summed, this single event will have a large impact on the outcome. If a Wilcoxon test were calculated, the data pair containing this epoch would get a top rank, but the size of the deviation is not taken into account. This can be an advantage or a disadvantage, depending on the situation and on the way a purported DMILS effect operates (single burst vs. continuous shift). Therefore, the decision of which epoch aggregation procedure is more suitable cannot be made theoretically but only by empirical data. This method was used in recent DMILS/remote staring studies as either a major evaluation strategy or an additional method for comparison reason (see, e.g., Wiseman & Smith, 1994).

Advantages

Similar to the PIS, an indicator of success for each session can easily be calculated and communicated to participants. This indicator is standardized by mean and standard deviation representing a good estimate of the session's outcome.

Disadvantages

Calculating two tests on the same data successively always leads to a decrease in power. Therefore, it should be avoided in DMILS research. Most of the DMILS/remote staring studies are underpowered anyway if one takes into account the estimated effect size of .25 (Schlitz & Braud, 1997).

The most important reason to avoid this procedure, however, is the fact that it violates one assumption of the test. The Wilcoxon test implies that pairs to be compared are dependent, but it assumes that the pairs generated as the basis of the comparison are themselves single, independent measurements. With all pairs of data to be tested stemming from the same participant this assumption of independence is violated. Therefore, this procedure is not appropriate for the data.

PAIRED *t* TEST

One can calculate the sum (or the mean) of all epochs of one session for each of the two conditions. That way, one obtains a pair of data for each session. Next, a paired *t* test over all sessions is calculated. This test has not been used in DMILS research so far.

Advantage

The paired *t* test has a high statistical power compared to other nonparametric tests. This test is the first method presented that has a powerful approach to the data with a higher chance of detecting DMILS effects.

Disadvantage

The *t* test assumes normal distribution of the differences of the data pairs and homogeneity of variances (Bortz, 1993). In some cases, a normal distribution of EDA data cannot be guaranteed; it is dependent on the parameters used. EDA generally shows very strong interindividual differences (Vossel, 1990). Hence, this test might not be appropriate as a prespecified evaluation method. In addition, it does not provide an outcome measure for a single session.

Some researchers have used a so-called *difference score* or *detect score* (Schlitz & LaBerge, 1997; Wiseman & Schlitz, 1997, 1999). To obtain it, a mean or sum of all epochs of both experimental conditions per session is calculated in the same way as for the paired *t* test described above. This data pair is then reduced to a single value by subtracting the values for each condition. Then a single mean *t* test assesses whether the empirical mean differs significantly from zero. This procedure is in fact the same as the paired *t* test. Testing data pairs with the paired *t* test, or subtracting them and using a single mean *t* test, yields exactly the same results, so advantages and disadvantages described for the paired *t* test also apply to this method.

WILCOXON SIGNED RANK TEST

If a paired *t* test cannot be applied because the distribution of the data is unknown, the nonparametric Wilcoxon signed rank test can be applied. So far this test has not been used in DMILS research. This Wilcoxon method differs from the one presented in the Two-Component Model section by the level of data to which it is applied. While the Wilcoxon test in the two-component model is applied to epoch data for calculating a session score, the Wilcoxon test in this section is applied to session data for the calculation of a *z* score for the whole experiment.

Advantage

The test requires only minimal distribution assumptions. It has a reasonable power that outperforms the PIS and the two-component model in most cases.

Disadvantage

This method cannot provide a single-session outcome measure.

RANDOMIZED PERMUTATION ANALYSIS

Blair and Karniski (1993) introduced a method for significance testing of waveform difference potentials. This method was developed for the significance testing of event-related potentials of different conditions in electroencephalography research. However, the procedure also fits DMILS data, because the basic characteristic of the data (i.e., many-data-points, few-subjects type of problem) are more or less the same. Dean Radin has adapted this procedure in an excellent way to DMILS data. We will not describe the detailed procedure here (see the Appendix of Radin et al.'s, 1998, article, for a full description). The basic idea of that strategy is to superimpose the data and to calculate a single outcome score for the whole experiment. Then the assignment of the experimental conditions for at least one session is exchanged completely, and the outcome score is calculated again. This procedure is repeated until all possible permutations of experimental conditions to the data have been calculated. The outcome values obtained by the permutation form an empirical distribution, and the ratio of number of outcome scores larger than the one obtained in the experiment to the number of all possible outcome scores by permutation give the true p value (Edgington, 1987).

Advantage

This is a 100% power test without any distribution assumptions. Therefore, it represents a perfect solution to most of the problems posed by the above-mentioned tests.

Disadvantage

For N sessions 2^N permutations have to be calculated. For example, in our pilot study $2^{26} = 67,108,864$ permutations would have had to be calculated. Radin, Machado, and Zangari (1998), however, have shown that the procedure can be shortened by an approximation as the p value converges rapidly toward the result of the full permutation. However, this test also does not provide a single-session outcome.

RESULTS

We have applied all methods described above (except the detect score) to the data of our pilot study. Figure 6 shows the results for the data based on the tonic SCL. A Kolmogorov-Smirnov goodness-of-fit test revealed p values of .21 for the differences between calm and activate means over all epochs (per session), indicating that the empirical distribution of the differences did not depart significantly from normality. Our random permutation analysis was programmed in C++ and run on an SGI Unix Workstation. The program was designed to stop the permutation process as soon as the true p value proved to be stable in the fourth decimal. Approximately 1,600,000 permutations had to be calculated.

Figure 7 depicts the results for the three parameters computed for the phasic component. The Kolmogorov-Smirnov goodness-of-fit test revealed p values of .97 (number of NS.SCR $>.015 \mu\text{S}$), .99 (number of NS.SCR $>.02 \mu\text{S}$), and .37 (sum of amplitudes) for the differences between activate and calm means. These values justify the application of a paired t test. The random permutation analysis was run on the same aggregated data pairs that also entered the paired t test and the Wilcoxon test.

Because our (psychophysiological) parameterization yields only small numbers for the parameters with large variances, the PIS method has only a very low power to detect the effect. All t scores are less than half of other methods' z scores. All other procedures yielded approximately the same results with only minor differences.

DISCUSSION

Our pilot study was conceptualized to test our equipment and to assess new methods in DMILS research. It yielded strong effects. The strong effects within the data gave us the chance to compare different approaches to the evaluation of DMILS data. Fortunately, we were able to sufficiently answer all research questions for which the study was designed.

We could demonstrate effects of approximately the same size in both components of EDA. This indicates that the effect is more of a global one rather than a very specific influence. The activation effect was to be found in the slow-changing SCL as well as in the amount and size of SCRs. We conclude that the activation intention effect is related to a more global physiological state reflecting overall arousal.

This finding fits well with those regarding the influence of irregularities in respiration. Some of the changes within the phasic EDA could be clearly linked to the pulmonary system. Our data indicate that EDA responses that are elicited by pulmonary irregularities are to be conceived of a substantial part of the DMILS effect. Therefore, the link between the pulmonary and electrodermal systems seems to be very important for

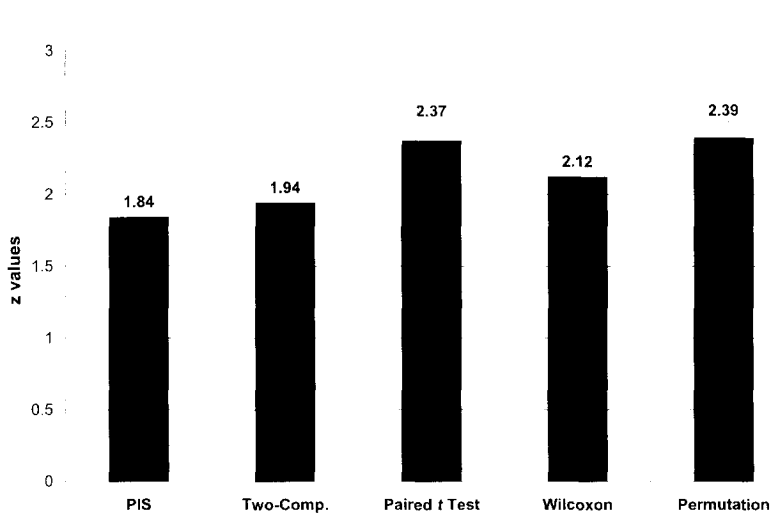


Figure 6. Comparison of activate and calm data taken from the tonic skin conductance level using different statistical procedures. For explanation of the different methods, see the text. As can be seen, the permutation procedure and the paired *t* test yielded the highest *z* scores, whereas the smallest *z* scores were observed for the traditional percentage influence scores (PISs). Two Comp. = two-component model.

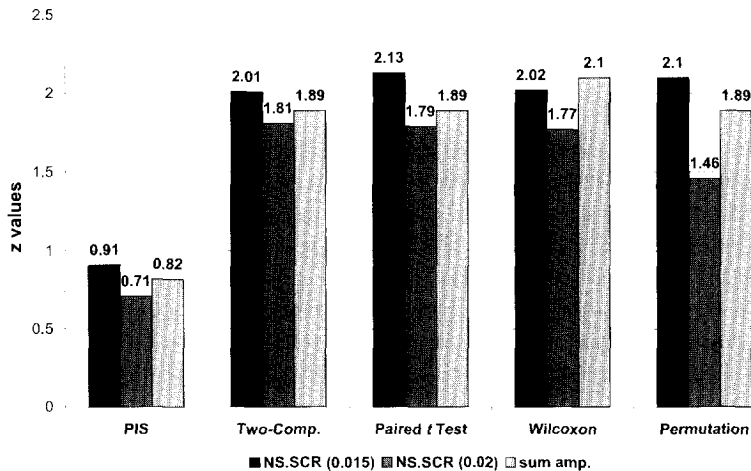


Figure 7. Comparison of activate and calm data for three different parameters (sum of amplitudes and number of nonspecific skin conductance responses [Ns.SCR] with minimum thresholds of either 0.015 μ S or 0.02 μ S) based on the phasic SCR using of different statistical procedures. For explanation of the different methods, see the text. Two Comp. = two-component model.

EDA–DMILS research. If this result proves to be true, it is not necessary to exclude respiration artifacts in DMILS data.

Our results also indicated that a psychophysiological way of parameterization of the phasic DMILS data was more suitable. Our data clearly revealed that our method of using the classical psychophysiological parameters (numbers of NS.SCRs, sum of amplitudes of NS.SCRs) outperformed the method traditionally applied in DMILS/remote staring research. If this is replicated in other studies, it can be concluded that all experiments using the phasic EDA might have underestimated the effect because of an inappropriate parameterization.

Furthermore, we can conclude that the effect we found acted specifically on the phasic responses as the effect increased for methods of parameterization that took into account only these responses (number of NS.SCR, sum of amplitude of NS.SCR) and left the flat area between them untouched. Taking smaller responses into account by lowering the threshold for responses down to $0.015 \mu\text{S}$ yielded slightly better results than a threshold of $0.02 \mu\text{S}$. This could be due to the fact that ψ acts more on small responses. On the other hand, the difference is too small ($z_{0.015} = 2.02$ vs. $z_{0.02} = 1.77$) for an interpretation, but a hypothesis for future research can be derived.

Regarding statistical procedures, the most important finding of our pilot study is the obvious inappropriateness of the PIS method. This conclusion applies to both the interpretation of PIS as a score of amount of influence expressed in a percentage value as well as to the significant testing based on these scores. We highly recommend abandoning this method for ψ research. Also, the other method (two-component model) used so far in DMILS research is not an appropriate procedure, because a basic assumption of the test is violated.

We favor the Wilcoxon signed rank test as the simplest method. It is easy to calculate and free of any distribution assumptions. This might be important, because EDA data from different participants cannot be guaranteed to have homogenous variances. If the difference scores between activate and calm epochs are normally distributed, the assumptions for a paired t test are fulfilled. This parametric testing procedure might result in a slight increase in statistical power. Full (100%) power can be reached by the permutation test, but this method is not so easily applied and tends to take much time with large session numbers. The program used here can be obtained from Stefan Schmidt.

We still see a need for further development of evaluation methods within the DMILS paradigm. The aim of those methods should be not a mere significance testing of a whole data set but rather a more detailed approach to the data.

CONCLUSION

The findings of this exploratory pilot study are consistent with previous findings in DMILS research. Our results indicate that EDA is the appropriate variable to map a global activation effect that affects several physiological systems. Earlier studies might have underestimated this effect by the use of inappropriate statistics and unorthodox parameterization methods.

These exploratory findings need to be replicated in future experiments. We have decided to record respiration activity as well as tonic and phasic EDA parameters in our future studies. We regard the number of NS.SCRs being larger than 0.015 μS as the most appropriate tonic parameter derived from the phasic EDA component. The findings regarding the impact of irregularities in respiration on DMILS results will be investigated in future data sets. The large field of research within EDA data calls for future studies addressing differential hypotheses toward the nature of the supposed effect.

We think that the DMILS paradigm may benefit from the approach described herein both as far as the detection of any ostensible effect is concerned as well as the acceptance of mainstream psychology.

REFERENCES

- BLAIR, R. C., & KARNISKI, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology*, **30**, 518–524.
- BOLLER, E., & SCHWEIZER, K. (1992). *Methoden der Generierung strukturierter Zufallszahlen. Forschungsberichte des Psychologischen Institutes der Albert-Ludwigs-Universität Freiburg i. Br. Nr. 79* [Methods for the generation of random numbers with different structures. Research report of the Psychological Institute of the University of Freiburg]. Freiburg, Germany: Universität, Psychologisches Institut.
- BORTZ, J. (1993). *Statistik für Sozialwissenschaftler* (4., vollst. überarb. Aufl.) [Statistics for the social sciences (4th revised edition)]. Berlin: Springer.
- BORTZ, J., & DÖRING, N. (1995). *Forschungsmethoden und Evaluation* [Methods of research and evaluation]. Berlin: Springer.
- BOUCSEIN, W. (1992). *Electrodermal activity*. New York: Plenum.
- BOUCSEIN, W. (1995). Die elektrodermale Aktivität als Emotionsindikator [Electrodermal activity as an indicator of emotions]. In G. Debus, G. Erdmann, & K. W. Kallus (Eds.), *Biopsychologie von Stress und emotionalen Reaktionen. Ansätze interdisziplinärer Forschung* (pp. 143–161). Göttingen, Germany: Hogrefe.

- BRAUD, W. G., & SCHLITZ, M. J. (1989). A methodology for objective study of transpersonal imagery. *Journal of Scientific Exploration*, **3**, 43-63.
- BRAUD, W. G., & SCHLITZ, M. J. (1991). Conscious interactions with remote biological systems: Anomalous intentionality effects. *Subtle Energies*, **2**, 1-46.
- BRAUD, W. G., SCHLITZ, M. J., COLLINS, J., & KLITCH, H. (1985). Further studies of the Bio-PK-effect: Feedback, blocking; specificity/generalality [Abstract]. In R. A. White & J. Solvvin (Eds.), *Research in parapsychology 1984* (pp. 45-48). Metuchen, NJ: Scarecrow Press.
- DELANOY, D. L., MORRIS, R. L., BRADY, C., & ROE, A. (1999). An EDA DMILS study exploring agent-receiver pairing. *Proceedings of Presented Papers: The Parapsychological Association 42nd Annual Convention*, 68-82.
- EDGINGTON, E. S. (1987). *Randomization tests*. New York: Marcel Dekker.
- FOWLES, D. C., CHRISTIE, M. J., EDELBERG, R., GRINGS, W. W., LYKKEN, D. T., & VENABLES, P. H. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology*, **18**, 232-239.
- LORIG, T. S., & SCHWARTZ, G. E. (1990). The pulmonary system. In J. T. Cacioppo & L. G. Tassinari (Eds.), *Principles of psychophysiology: Physical, social and inferential elements* (pp. 580-598). Cambridge, England: Cambridge University Press.
- RADIN, D. I., MACHADO, F. R., & ZANGARI, W. (1998). Effects of distant healing intention through time and space: Two exploratory studies. *Proceedings of Presented Papers: The Parapsychological Association 41st Annual Convention*, 143-161.
- REBMAN, J. M., RADIN, D. I., HAPKE, R. A., & GAUGHAN, K. Z. (1996). Remote influence of the autonomic nervous system by a ritual healing technique. *Proceedings of Presented Papers: The Parapsychological Association 39th Annual Convention*, 133-147.
- SCHAEFER, F. (1999a). EDR PARA (Version 3.71) [Computer Software]. Wuppertal, Germany: Author.
- SCHAEFER, F. (1999b). EDR Select (Version 2.1) [Computer Software]. Wuppertal, Germany: Author.
- SCHAEFER, F., & BOUCSEIN, W. (2000). Comparison of electrodermal constant voltage and constant current recording techniques using the phase angle between the alternating voltage and current. *Psychophysiology*, **37**, 85-91.
- SCHLITZ, M. J., & BRAUD, W. G. (1997). Distant intentionality and healing: Assessing the evidence. *Alternative Therapies in Health and Medicine*, **3**, 62-73.
- SCHLITZ, M. J., & LABERGE, S. (1997). Cover observation increases skin conductance in subjects unaware of when they are being observed: A replication. *Journal of Parapsychology*, **61**, 185-196.

- SCHMIDT, S., & WALACH, H. (2000). Electrodermal activity (EDA)—State-of-the-art measurement and techniques for parapsychological purposes. *Journal of Parapsychology*, **64**, 139–163.
- SPSS FOR WINDOWS 9.0 [COMPUTER SOFTWARE]. (1998). Chicago, IL: SPSS Inc.
- VENABLES, P. H., & CHRISTIE, M. J. (1980). Electrodermal activity. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 3–67). Chichester, England: Wiley.
- VOSSEL, G. (1990). *Elektrodermale Labilität. Ein Beitrag zur Differentiellen Psychophysiology* [Electrodermal lability. A contribution for differential psychophysiology]. Göttingen, Germany: Hogrefe.
- WEZELMAN, R., RADIN, D. I., REBMAN, J. M., & STEVENS, P. R. (1996). An experimental test of magical healing rituals in mental influence of remote human physiology. *Proceedings of Presented Papers: The Parapsychological Association 39th Annual Convention*, 1–12.
- WISEMAN, R., & SCHLITZ, M. J. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology*, **61**, 197–207.
- WISEMAN, R., & SCHLITZ, M. J. (1999). Experimenter effects and the remote detection of staring: An attempted replication. *Proceedings of Presented Papers: The Parapsychological Association 42nd Annual Convention*, 471–479.
- WISEMAN, R., & SMITH, M. D. (1994). A further look at the detection of unseen gaze. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, 465–478.

Institute of Environmental Medicine and Hospital Epidemiology (IUK)
Universitätsklinikum Freiburg
Hugstetter Str. 55
D-79106 Freiburg
Germany
sschmidt@ukl.uni-freiburg.de

Institut für Grenzgebiete der Psychologie und Psychohygiene e. V.
Wilhelmstr. 3a
D-79098 Freiburg
Germany
raischnei@aol.com.