

## Experimenter Effects in Laboratory Tests of ESP and PK Using a Common Protocol

CHRIS A. ROE AND RUSSELL DAVEY

*Division of Psychology, The University of Northampton, UK*

PAUL STEVENS

*Koestler Parapsychology Unit, University of Edinburgh, UK*

**Abstract**— This paper describes the fourth in a series of studies that explore the relationship between ESP and PK performance by testing for both using a common protocol so as to control for expectancy effects and experimental artifacts. Following earlier work we were particularly concerned to look for evidence of experimenter effects. Forty participants completed a computer-based greyhound racing game. Races occurred in two blocks of 12, with one block presented as an ESP task and the other as a PK task, though in fact each block included equal numbers of ESP and PK trials presented in random order. Roe and Davey each served as experimenter for 20 sessions and, after briefing each participant, would rate the interaction for warmth, spontaneity and positivity. Performance was non-significantly better than chance overall, but was significantly so for disguised ESP trials ( $p = 0.011$ ). As predicted, participants who had been briefed by Roe performed better overall than those briefed by Davey; suggestively so overall ( $p = 0.085$ ) and significantly so for disguised ESP ( $p = 0.002$ ). Some interaction measures gave promising correlations with task performance, particularly the experimenter's confidence of success ( $r_s = -.431, p = 0.007$ ).

**Keywords:** ESP—PK—experimenter effects

### Introduction

In recent work (Roe et al., 2003a, 2004, 2005) we have been concerned to address the question of whether ESP and PK functioning are sufficiently distinct to merit separate terms. Earlier research that had considered this issue is difficult to interpret because the method of testing for ESP is typically quite different from that for PK so that any apparent differences in the preferred conditions of the phenomena may be artificial (Schmeidler, 1988). We developed a new protocol using a computer game interface that allowed both phenomena to be tested for within a standardised context. In the game, random number generator (RNG) and pseudorandom data are sampled to determine the movements of six greyhounds from the left to the right of the screen, simulating a race (see Figure 1). The program monitors progress and registers the order in which the dogs

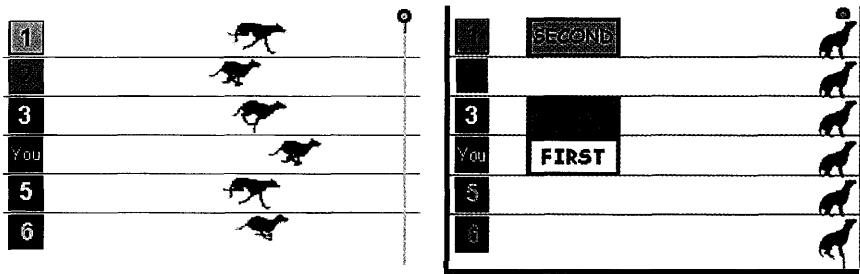


Fig. 1. Screenshots of greyhound race and of race finish.

cross the finishing line. In the ESP condition a race had been run 'silently' so that the outcome was 'known' to the computer. Participants were informed that their task was simply to select one dog from among the six that they felt had performed best on that trial. They then watched a replay of the race and the result was confirmed. In the PK condition the race would be run in real time with the movements of their pre-selected greyhound determined by an RNG. Participants were informed that their task was to attempt to influence the RNG and thus enable their greyhound to succeed. The program consisted of a block of 12 races that ostensibly were all testing for ESP and a further block of 12 testing for PK. However, half of the trials that appeared to be tests of ESP in fact were of PK and vice versa in order to differentiate between characteristics of the phenomenon and participants' expectancies concerning that phenomenon.

The results from our first three studies have been somewhat disappointing, with overall performance at chance levels for both ESP and PK trials, and for true and disguised trials. One potential contributory factor to poor performance that we have not considered is that variables associated with the experimenter may have had an inhibiting effect. White (1977, p. 273), for example, has noted that "the experimenter has been a neglected variable in parapsychological research ... [yet] ... there could hardly be a more significant area of investigation than the role of the experimenter". Rhine and Pratt (1957) have characterised the experimenter as having to be able to provide "the psychological conditions under which psi can operate" (p. 131, cited in White, 1977, p. 274), and Gardner Murphy (1949) suggested that there is no such thing as a gifted participant as such, but rather how well a participant scores on a psi task depends on the person who does the testing and the nature of the experimental conditions. These experienced researchers seem to share the view that the experimenter plays a crucial role in encouraging or inhibiting psi in the laboratory (see Smith, 2003a, for an updated consideration of this issue).

Within parapsychology there seems to be a common belief that some experimenters are psi-conducive, whereas others appear to be psi-inhibitory (cf. Irwin, 1999; Smith, 2003a). Indeed there is quite clear evidence of an experimenter

effect leading to differential performance on psi tasks in circumstances where all other conditions (and even in some instances the participants themselves) are common to both. For example, Van Busschbach (1956) describes a study in which Rhea White and Margaret Anderson tested schoolchildren in different classes but of the same age and in the same schools. The latter oversaw above-chance scoring ( $p = 0.002$ ) whereas the former's participants performed at chance, with the difference between experimenters significant ( $p = 0.02$ ). Similarly, Nicol and Humphrey (1953) found striking differences in the results of the same participants when each of the authors served as experimenter using the same test under similar conditions, and Bednarz and Verrier (1969) reported a similar finding, albeit with overall negative scoring. In an important series of studies intended to explore experimenter effects within the staring detection paradigm, a prominent parapsychologist who had been consistently successful in eliciting psi effects, Marilyn Schlitz, and a prominent sceptic who had consistently been unsuccessful, Richard Wiseman, collaborated to run sessions using the same equipment and participant pool. The first two studies (Wiseman & Schlitz, 1997, 1999) replicated their earlier pattern of performance, with Schlitz's participants scoring significantly better than Wiseman's, to give a clear indication that the experimenter's role may be a pivotal one (but note that a third study, cited by Watt et al., 2005, failed to replicate this difference).

One possible explanation of the effect is that the experimenter's personality, behaviour and enthusiasm may indirectly influence the results of a psi study by motivating participants or providing them with clues that provide further information about the nature of the study and about the experimenter's hopes or expectations. These demand characteristics may affect the subsequent behaviour of the participants and thus the results of the study itself (Harris & Rosenthal, 1985; Rosenthal, 1966; see also White, 1977). For example, in terms of their communication skills, there is some evidence that, at least in a non-experimental context, psi-conducive and psi-inhibitory experimenters differ in how their body language is perceived by observers. Schmeidler and Maher (1981) video taped researchers as they gave talks and fielded questions at an academic conference. Five who were regarded as psi-conducive were matched for relative age and physical features with five who were regarded as psi-inhibitory, and their video footage was shown—with sound levels too low to hear—to independent judges, who rated the researchers along a number of dimensions. Psi-conducive experimenters were considered to be, *inter alia*, more flexible, enthusiastic, friendly, likeable and warm and less tense, irritable and cold. Of course, it is possible that the way that researchers present themselves at a parapsychological conference is affected by the success or otherwise of the research they are describing, and of their own sense of how they might generally be perceived by their peers, though this suggestion is dismissed by Schmeidler and Maher (1981)—rather too casually in our view. In a more direct assessment, Honorton et al. (1975) had two experimenters who interacted either in a positive manner (friendly, casual and

supportive) in which time was taken to establish rapport with the participant, or in a negative manner (abrupt, formal and unfriendly) in which they went quickly into the task. The positive treatment gave significantly higher scores than the negative. However, when Schneider et al. (2000) manipulated the experimenter's interactional style in an EDA-DMILS study that seems analogous to warmth ('personal' versus 'neutral'), they found no difference between the conditions; this could be a function of the different demands of this type of study in comparison to the ESP studies that predominate in the experimenter effects literature.

There is also evidence to suggest that the beliefs of the experimenter might influence the outcome of the study. Smith (2003b) surveyed researchers who had served as experimenters in at least one published parapsychological experiment, measuring attitude and personality variables as well as asking them to rate other researchers in terms of the degree to which they were psi-conducive. By combining these latter ratings across respondents, Smith was able to gauge the general perception of each researcher's success in eliciting psi in the laboratory, and found that this was significantly correlated with their self-ratings of abstract belief in both ESP and PK and also in their own psi ability. Of course, we cannot infer from these findings that greater belief leads to experimental success, since it is likely that one's beliefs will be affected by exposure to one's own positive or negative research results. However, when Sharp and Clark (1937) employed several experimenters within the same study who differed in their prior attitude towards psi, they obtained results that did seem to relate experimenters' attitudes to participants' subsequent hit rates; Sharp (mean hit rate = 5.36, where mean chance expectation [MCE] = 5.00) and Davidson (5.88) were positive towards the existence of psi, whereas Berger (4.86) was uncommitted, and Myers (4.30) was sceptical. However, typically for a quasi-experiment, it is difficult to know whether these experimenters may have differed from one another in other ways that might have influenced participant scoring, for example the more sceptical experimenters may have been less friendly, older, etc. Parker (1975) more directly manipulated expectancy among six 'experimenters' (student data collectors) so that they presented as strong believers or strong disbelievers. Although overall scoring was null, there was a significant difference in performance between the experimenter groups, even though they in fact were testing the same sender-receiver pairs. Watt and Ramakers (2003) adopted this protocol in exploring experimenter effects within a remote facilitation of attention task. They trained nine believers and five disbelievers to serve as experimenters for randomly allocated participant pairs. They reported a significant overall remote facilitation effect that they attributed to the believer-experimenter condition; participants working with believer-experimenters scored significantly better than those working with disbeliever-experimenters (who performed at chance levels). Two earlier remote facilitation of attention studies (Watt & Baker, 2002; Watt & Brady, 2002) had failed to find evidence of experimenter effects, but the condition manipulation in these cases relied on role-play and instructional set respectively, and so may not have been as effective.

Alternatively it may be that different outcomes reported by different experimenters are explicable in terms of some form of parapsychological experimenter effect—where the results are partially dependent on the experimenter's own psi ability (see Kennedy & Taddonio, 1976; Palmer, 1997; Schmeidler, 1997; White, 1976). However, on grounds of parsimony this study will primarily focus on psychosocial explanations.

Reflecting on the previous studies in this series, we should note that the researcher responsible for all interactions with those participants was the second author (R.D.). Although R.D. has a Bachelor's degree in Psychology and has previously conducted a parapsychological study for his dissertation, he would nevertheless be considered a novice experimenter, whereas the first author (C.A.R.) has been involved in a number of previous studies that have reported significant effects (e.g., Roe, 1996; Roe et al., 2003b) and may simply be more practiced at engendering a psi-conducive atmosphere. It could be that if a more experienced researcher had interacted with participants in these earlier studies, then a more positive outcome might have occurred. Secondly, although involved in the later stages of design of these studies, R.D. was not involved at the project's inception and may not feel the same degree of 'ownership' of the project that C.A.R. would feel through having been responsible for the seed idea, conducted background literature research, written funding proposals and so on. On these grounds our primary hypothesis was that participants tested by C.A.R. would perform significantly better than those tested by R.D.

We were also interested in the interplay between the researcher's personality and mood and their interaction with the participant as a vehicle for any experimenter effect. We suspected that these factors might not remain constant within any individual but rather might be prone to fluctuate from time to time in response to a legion of subtle and not so subtle situational variables. We therefore planned to have the experimenter complete an assessment of key aspects of the interaction for each session to explore whether these might predict successful trials, including (after Schmeidler & Maher, 1981) dimensions of flexibility (which we reframed as spontaneity), warmth and tension and (after Woodruff & Dale, 1950) degree of rapport with the participant.

## Method

### *Design*

This study incorporated a  $2 \times 2$  mixed design looking at the effects of task type (ESP versus PK) and experimenter (C.A.R. or R.D.). Participants were allocated randomly to an experimenter.<sup>1</sup> The former involved repeated measures comparisons, while the latter involved independent samples comparisons. The primary outcome measure was pre-specified to be the weighted sum of ranks of

finishing positions. We also intended to conduct exploratory correlational analyses to determine whether task performance in the four conditions covaried systematically with experimenter mood and interaction variables. All analyses were planned to be non-parametric.

### *Materials and Apparatus*

A participant information form (PIF) was constructed which asked about basic biographical and contact details. The PIF is a generic form that also includes various questions (e.g., about paranormal belief and experience) that were not planned to be a focus of this report. Copies of the PIF are available on request from C.A.R. A seven-item interaction questionnaire was devised specially for this study (a copy is included as an Appendix). Items were included to gauge the experimenter's mood at the time of their interaction with the participant, as well as their rating of the warmth, spontaneity and positivity of the interaction (after Schmeidler & Maher, 1981; Woodruff & Dale, 1950) and confidence of success (after Parker, 1975).

The program consists of 24 races, run in two blocks of 12 that ostensibly are either tests of ESP or PK. In fact within each block half the trials are of ESP and half are of PK, presented in random order. ESP and PK trials are distinguishable in terms of the degree to which participants are free to choose which greyhound will be theirs, whether the race is run in real time or the outcome is already 'known' to the program and how theoretically susceptible the source of randomness is to any possible PK influence.

Practically, the four conditions are distinguishable as follows:

1. **True ESP trials:** The greyhound race was run silently before the trial using pre-recorded random data. The outcome was recorded on PC hard disk so that it was theoretically available to participants before they freely selected their greyhound. The race was subsequently 'replayed' on screen.
2. **True PK trials:** The race was run in real time using 'live' RNG data. Participants were allocated one of the six dogs using a file with data from <http://random.org> (so that there is no opportunity for them to make their own selection in a manner that could be informed by ESP).
3. **Disguised ESP trials:** Again the trial is actually pre-run and outcome 'known' to the PC. Participants 'select' their dog by the timing of their space bar keypress, allowing for an interpretation in terms of Decision Augmentation Theory. Although participants believe they are watching the race in real time it is in fact a replay.
4. **Disguised PK trials:** Participants apparently 'select' one of the six dogs as for the true ESP condition, but in fact their greyhound is chosen for them by the computer using a pseudorandom data file. Where these choices differ, the program switches the data so that the movement of the participant's chosen greyhound is determined by the data originally intended for the computer's chosen greyhound and vice versa (so that,

effectively, greyhound 2 is running in lane 5 and greyhound 5 in lane 2, for example). The trial continues as for the true PK condition.

Pre-recorded data are used for ESP trials rather than real-time data from the RNG as these should be less open to any PK influence. True random data were collected using an Orion Random Number Generator, which consists of two independent analogue Zener diode-based noise sources. Both signals are converted into random bit streams, combined (via a NAND gate) and subsequently transmitted to the computer in the form of bytes via the RS-232 port. (For more information, visit <http://www.randomnumbergenerator.nl/rng/home.html>.) 'Random' data for selected dogs in ESP trials were drawn from a single data file generated before the study began by taking true-random atmospheric noise data available from <http://www.random.org/nform.html>. 'Pseudorandom' data for non-selected race dogs in ESP trials were generated using the QBASIC RND function. For both forms of pseudorandom data, the integer was converted to binary format and the 1s were added up to calculate by how much a dog's position should be advanced each time, so that over successive iterations some greyhounds move closer to the finish than others. The program monitors progress and notes the order in which the dogs cross the finishing line. The program continues until all six dogs have completed the course.

### *Participants*

Forty people participated in this study: 21 men and 19 women, with a mean age of 31.4 years ( $SD = 13.6$ , *Median* = 24). Participants were drawn from an opportunity sample from the Northampton area. Relatively few were undergraduate students at The University of Northampton.

### *Procedure*

Prior to the session, participants were given the PIF to take away and complete. They were greeted by the first or second author (C.A.R. or R.D.) who acted as experimenter: In some cases, participants had not completed the measure (e.g., if they had questions about certain items) in which case they were given time prior to their trial to complete the form. Participants next completed the state form of Spielberger's (1983) State-Trait anxiety inventory. The experimenter sought to make the participant feel relaxed and comfortable, engaging in casual conversation as well as explaining the nature of the task and answering any questions they might have.

Participants were then escorted by C.A.R. or R.D. into a research cubicle containing a PC with the program ready to begin. The program autoran and presented participants with a series of 24 races in two blocks of 12, taking approximately 12 minutes to complete. One block was labelled as 'gambler' races and were ostensibly ESP trials. Here participants saw the onscreen briefing:

For the next 12 trials we'd like you to play the role of a gambler who has a free hand to choose which dog to select. In this session the races will already have been run by the computer but not yet have been played out. Your task is to use ESP to identify which of the 6 dogs won the race. Once you've made your choice you'll see a replay of the race on screen.

Prior to each gambler race, participants were prompted to enter a number from 1 to 6 corresponding to their choice of dog for the forthcoming 'replay'. A second block was labelled as 'owner' races and consisted of ostensible PK trials. Here the onscreen briefing was:

For the next trials you will play the role of an owner whose greyhounds are entered in a series of races. Your dog will be pointed out at the beginning of each race, and its speed will be determined by a random number generator in the computer. Your task is to try to use PK to influence the RNG so that your preselected dog wins the race. You'll see the race in real time so you get feedback on how well you're doing.

Prior to each owner race, participants were asked to press the space bar to start the race. All participants completed both blocks with the order of completion counterbalanced across participants. Within each block, half the trials were as given in the briefing (e.g., tested for ESP in the gambler block), but half were not (e.g., tested for PK in the gambler block) to gauge the effect of expectation on performance. Prize money is used as a simple weighted score based on finishing position (100 virtual pounds for first, £50 for second, £25 for third, no prize money for the other placings). After a series of races the participant amasses an amount of overall prize money.

The experimenter remained outside the research cubicle during trials but was available should assistance be required. During this time (i.e., before the outcome was known) the experimenter completed the interaction questionnaire. Once the participant had completed the first block of 12 races, the experimenter rejoined them to discuss how they were getting on and to engage in casual conversation so as to help reduce any fatigue and/or boredom and to reinforce the experimenter-participant relationship. They again waited outside while the participant completed the second block of trials. After the program had finished, the experimenter debriefed participants, describing the nature of the four conditions within the task and explaining the need to disguise certain aspects of it. Given the mild deception involved, great pains were taken to ensure that participants were satisfied of the need for the study to be designed as it was and to be sure that they were happy for their data to be included in analysis. None of the participants asked to withdraw.

## Results

The planned outcome measure here is the sum of ranks (SOR) of finishing positions for participants' greyhounds in computer races, but to get a sense of whether overall performance was above MCE we shall firstly consider the

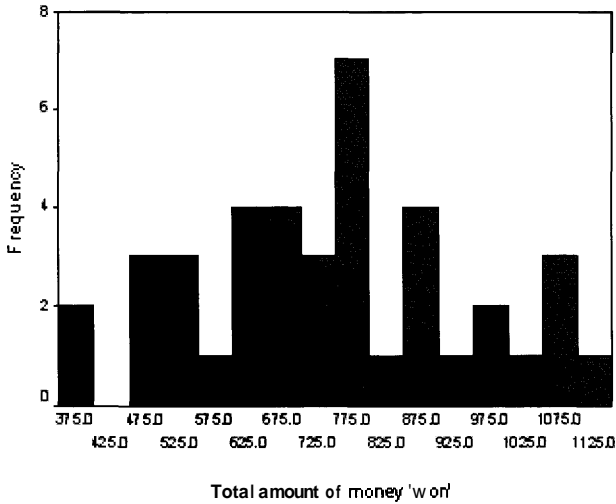


Fig. 2. Frequency histogram of prize money 'won' by participants.

overall amount won by each participant. The greater the success at the task, the greater the amount of prize money that will have been won. If chance alone is operating then a participant will typically have won prize money of £700. We can see from Figure 2 that the distribution of scores is relatively normal and peaks slightly to the right of the theoretical midpoint; the average prize money is non-significantly above this ( $M = £730.0$ ,  $SD = £198.8$ ; Wilcoxon  $Z = -.838$ ,  $p = 0.402$ , 2-tailed).

The distribution of ranks for each of the four conditions is given in Table 1. We can see that in terms of overall scoring, results in this study are somewhat better than previously. The overall sum of ranks for target dogs is below the MCE of 840 in two of the four conditions, significantly so in the case of disguised ESP (note that lower sums of ranks indicate better performance in races). Performance in the two PK conditions closely approximates MCE, and

TABLE 1  
Sum of Ranks (SOR) for Greyhound Finishing Position

Condition	Finishing position						SOR	Z score	Effect size (r)
	1	2	3	4	5	6			
MCE	40	40	40	40	40	40	840		
True PK	43	46	31	31	41	48	845	.170	.011
Disguised PK	44	34	43	37	44	38	837	-.094	-.006
True ESP	38	30	48	40	34	50	872	1.191	.077
Disguised ESP	46	50	42	34	37	31	779	-2.287	-.148
Total	171	160	164	142	156	167	3333	-.501	-.016

TABLE 2  
Mean Sums of Ranks (and Standard Deviations) for Informed and Uninformed Participants  
for the Four Conditions

	True ESP	Disguised ESP	True PK	Disguised PK	Overall
R.D. trials	21.80 (4.76)	21.35 (3.66)	20.75 (4.01)	21.95 (3.86)	85.85 (8.98)
C.A.R. trials	21.80 (4.29)	17.60 (3.33)	21.50 (3.91)	19.90 (4.53)	80.80 (9.07)

for true ESP is non-significantly worse than chance expectation. Perhaps surprisingly, performance in the disguised ESP condition was significantly better than for the true ESP condition (reminiscent of Roe et al., 2005), although there is no overall difference in performance across the conditions (Friedman's  $\chi^2 = 5.52$ ,  $p = 0.20$ ).

Our principal interest in this study was to see whether there might be evidence of an experimenter effect, which may offer an explanation for previous null results (Roe et al., 2003a, 2004, 2005). The experimenter in those studies was R.D., who is a relative novice researcher, and it was hypothesised that C.A.R. might be more successful, especially given that the study was originally designed by him and that he may feel more 'ownership' over it. We can see in Table 2 that the overall performance of C.A.R.'s participants was superior to that for R.D., and this difference was suggestive overall,  $F(1,38) = 3.13$ ,  $p = 0.085$ . Comparing participants' performance across the four psi-task conditions (ESP and PK, either informed or disguised) suggests that there are differences in success across conditions that approach significance,  $F(3,114) = 2.31$ ,  $p = 0.080$ . Interestingly, the experimenter difference is more marked for the conditions that included an element of deception than those that did not (for the interaction between experimenter and psi condition,  $F(3,114) = 2.50$ ,  $p = 0.063$ ).

To further understand the nature of any possible experimenter effect, the relationships between experimenter ratings on the interaction questionnaire and task performance are given in Table 3. Given the exploratory nature of these analyses, there has been no correction for the increased likelihood of committing a Type I error resulting from multiple analyses. Were we to adopt the relatively stringent Bonferroni correction, the alpha level for significance would become  $p = 0.014$ , which for a study of this power would require very large effect sizes to remain significant. Rather than risk overlooking interesting associations, we leave the analysis uncorrected here but await confirmation of the effects in subsequent replications before we regard them as evidential. This caveat notwithstanding, we can see that there is a suggestive tendency for task performance to improve as the experimenter rates themselves as more relaxed. This seems particularly so for conditions that entailed deception. A similar pattern is evident with ratings of the orientation of the interaction, with greater positivity









