PSI ON THE WEB:
TWO REPLICATION STUDIES

BY FIONA STEINKAMP

ABSTRACT: Two studies were conducted on the World Wide Web to replicate the author's previous findings (Steinkamp, 1998). The first study examined whether participants who reported a positive reaction to filling out a questionnaire subsequently fared better on the psi task. Despite the higher power, the experiment failed to provide significant results. The second study investigated whether first trials could reliably perform better than subsequent trials. Again, despite higher power, the experiment failed to provide significant results. It is concluded that the initial work provided an inflated estimate of effect size. Although no significant results were obtained, further experimentation may still be useful.

Radin was one of the first experimenters to use networked computers to perform a psi experiment (Johnson, 1977). This was followed in 1984 by Tedder, who conducted a remote computerised parapsychological study in which selected participants accessed a site from internationally distant locations to take part in the experiment at a time of their choosing. However, Web experimentation in parapsychology using a self-selected population was initiated by Bierman (1995; Bierman & Wezelman, 1996a, 1996b) and has since been taken up by a number of researchers (e.g., McDermott, 1997; Rebman, Radin, & Stevens, 1996; Steinkamp, 1998; Stevens, 1997). Although Web experimentation is still very much in its infancy, if it can be used successfully, it has the potential to revolutionize psi research.

Once programmed, experiments on the Web can quickly accrue a large number of trials, thus enabling researchers to conduct studies with high power to detect more easily what is presumably a small effect. Moreover, the automated nature of such experiments should enable others to replicate them with comparative ease and without having to dedicate too much time to them. Of course, if Web experimentation is generally unsuccessful, it may serve as an indication that psi is not possible, at least under the conditions that the World Wide Web provides. Nevertheless, with the sheer diversity of ways in which studies can be conducted over the Web (e.g., allowing participants to perform only a few trials at one time, changing the amount of feedback, pre-selecting participants, and selecting forced-choice or free-response designs), it is too early to make any firm conclusions either way.

I would like to thank the two anonymous referees for their helpful comments.
A previous set of exploratory studies (Steinkamp, 1998) indicated that participants may perform particularly well on their first trials and that if participants fill out a questionnaire with a set of simple arithmetical questions and report a positive attitude to completing this questionnaire, their subsequent psi performance will be good. The studies reported here aimed to replicate these findings. The previous experiments had very low power, and so the number of trials was increased considerably for the subsequent studies reported here.

This article is divided into two parts. The first part provides details of the replication questionnaire studies. The second part describes an attempted replication of the comparison between performance of first and subsequent trials.

**Questionnaires and Psi Performance: A Direct Replication**

The previously published experiment (Steinkamp, 1998) compared psi performance after participants had responded to two different types of questionnaire: One ("mathematical") questionnaire consisted of a series of six simple arithmetical sums for participants to complete (e.g., 64/4), and the other ("intuitive") questionnaire comprised a number of simple verbal questions (e.g., “Do you consider yourself to be creative?”). The present study used the same two types of questionnaire, but with additional logical improvements. For example, this time both questionnaires had the same background throughout rather than using a more colorful one for the intuitive questionnaire. This previous design with a more colorful background on the intuitive questionnaire potentially made comparison between the two conditions difficult because it was unclear what influence, if any, a colored background had on participants’ motivation and performance. Similarly, whereas previously questions in the intuitive questionnaire were answered in a variety of different ways (e.g., input boxes, radio buttons allowing only one or more answers, or selection boxes in which participants opt for one or out of a number of choices) and those in the mathematical questionnaire used only input boxes, in this replication study both questionnaires used additional input boxes, in this replication study both questionnaires used input boxes.

Each questionnaire had eight questions. The questions were designed to require only one- or two-word responses from the participants to keep, as far as possible, the error of input similar for both questionnaires. Nevertheless, some participants clearly liked to expand on some of the verbal questions. There was no input limit to any question on either questionnaire.

On completing the questionnaire, participants had to indicate with a radio button allowing only one of two responses (positive or negative) how they had felt about having to answer the questionnaire. Participants then submitted their questionnaire, and the computer checked that the participant had answered all questions. Participants were then taken to the psi experiment.

For the psi task, the participant was shown pictures of six identical doors. The participant had to guess which one of the six doors the computer would later select to open. Thus the probability of a hit was 1/6, and the experimental design was a precognitive one. Once participants selected a door and submitted their guess, the computer immediately randomly selected a number between 1 and 6 using the clock to obtain the seed number. This number was then compared with the number of the door that the participant had selected. If participants guessed correctly, they saw an open door with streamers coming out of it along with some congratulatory text. If their guess was wrong, the door remained closed with the words "no easy" as feedback. Feedback was given as soon as the computer had selected the target.
On each trial, the program updated the total number of trials conducted so far after both questionnaire types, as well as the number of trials and hits following positive and negative responses to each questionnaire. The results were automatically stored on the KPU server. Although each participant was allowed only one trial, some participants may have taken part more than once if their access was through a server that does not retain the same location ID or if they moved to another computer.

A total of 2,500 trials was preplanned for each questionnaire type, making a total of 5,000 trials. This is almost 14 times more trials than in the 1998 study, which had only 180 trials per condition, and thus should suffice to produce a significant effect if the results from the previous study were reliable. Chi-square tests were preplanned to test the difference between positive and negative attitudes for each questionnaire, with the test being one-tailed for the numerical questionnaire, in the light of the results from the article published before and a two-tailed test for the verbal questionnaire because this was a new, more exploratory test. The overall results of the 5,000 trials were also planned to be analysed with a conservative two-tailed test. The effect size used is $\pi$, which can range from 0 to 1 (Rosenthal & Rubin, 1989). A null effect is .5, a $p$-value at a direction is above .50, and a negative direction is below .50. Trials were accumulated from January 1999 to February 2000. No attempt was made to compensate for response bias. To this extent, the analyses conducted are conservative.

**Results**

The results on this larger scale test failed to replicate those found before. Although there was still a nonsignificant tendency for those reporting a positive attitude to the numerical questionnaire to fare better on the subsequent psi task than those registering a negative attitude to it, performance was nevertheless consistent in the psi-missing direction. The details are provided in Table 1. The results do not even approach significance, $\chi^2(2, N=2500) = 0.63, p = .43$, one-tailed.

**Table 1**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Positive attitude</th>
<th>Negative attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>215</td>
<td>185</td>
</tr>
<tr>
<td>Misses</td>
<td>1,119</td>
<td>981</td>
</tr>
<tr>
<td>Total</td>
<td>1,334</td>
<td>1,166</td>
</tr>
<tr>
<td>% hit rate (MCE 16.67%)</td>
<td>16.12%</td>
<td>15.87%</td>
</tr>
</tbody>
</table>

Similarly disappointing results were found with the verbal questionnaire. The difference in performance from those reporting a positive attitude and those registering a negative one did not even approach significance, although, again, those who felt positively about the questionnaire performed nonsignificantly better, $\chi^2(2, N=2500) = 0.054, p = .81$, two-tailed. The results are presented in Table 2.

**Table 2**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Positive attitude</th>
<th>Negative attitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>349</td>
<td>79</td>
</tr>
<tr>
<td>Misses</td>
<td>1,699</td>
<td>379</td>
</tr>
<tr>
<td>Total</td>
<td>2,042</td>
<td>458</td>
</tr>
<tr>
<td>% hit rate (MCE 16.67%)</td>
<td>17%</td>
<td>17.25%</td>
</tr>
</tbody>
</table>

In addition, there was no evidence of psi with only 822 hits in the total 5,000 trials (16.44% hit rate) when 833 would be expected by chance ($z = -0.43, \tau = 0.496, p = .67$).

**Conclusion**

These results indicate that the findings from the 1998 exploratory study probably yielded an inflated effect size due to sampling error. On the basis of the previous effect size $\epsilon (\sqrt{N})$ of .18, the present study should have needed only 325 trials to obtain a significant effect at the 95% confidence level. In this respect, the present study was not underpowered. However, if the true effect size is similar to that found in other analyses—namely, $\epsilon = .01$ (e.g., Honorton & Ferrari 1989)—a post hoc power analysis reveals that the present study had a power of only .17. It is unlikely that the reason for the reduction in effect size is due to a change in the participant population, for there is no reason to believe that the KPU Web site’s readership has altered significantly. Nevertheless, the continued better, if nonsignificant, performance on those reacting positively to the relevant questionnaire may bear further investigation using higher powered studies.

**Investigating the Superior Performance of First Trials: A Replication Study**

The previous study (Steinkamp 1998), using the same six-door design as above, appeared to produce a relatively consistent pattern of superior performance on the initial trial taken by a participant on any one day. This was an interesting and potentially important finding. This experiment therefore attempted to replicate this result.

For the present study, the experiment was changed in the hope of making it more fun and engaging for participants. The Web site was made more colorful, and the text was written in a humorous way. When participants agreed to take part by clicking on the appropriate button,
their computer IP address was tracked and stored in a separate file, as in the experiment reported above. However, this time participants were allowed up to five consecutive trials in one day. Once participants entered the experiment, they were taken to a picture of "Freaky Fifi," who had green hair and was gagged. Below this picture was a row of three pictures of the back of Freaky Fifi’s head where you could see the knot of the handkerchief that was tied round her mouth. Participants were invited to predict which of the three pictures would later be selected to allow Freaky Fifi to be freed. Thus, the design was precognitive.

After participants had indicated their choice of one of the three pictures, the pseudo-random number generator (PRNG) on the server randomly selected the target cord in the same way as in the experiment reported above. This number was compared with the number of the picture the participant had chosen. If the participant correctly predicted which cord would be selected, they were taken to a Web page showing Freaky Fifi, un gagged, and smoking a cigarette. The computer also randomly selected 1 of 60 pieces of humorous “advice” to show the participant with the words “Freaky Fifi says ‘Thank you. My advice is [advice inserted here]’.”

Examples of advice are “Do not take advice from computer programs” or “You are on an winning streak.” By incorporating different, random pieces of advice on each successful trial, it was hoped that participants would be motivated to find out what the next piece of advice would be.

If participants had not chosen the correct target, they were taken to the same picture of the back of Freaky Fifi’s head, gagged, as they saw at the beginning of the trial, with the words “Oh no! Oh no! Freaky Fifi is still tied up!”

All of the participants were shown a button that enabled them to have another go if they wished. If they tried to do more than five trials, they were told that Freaky Fifi was tired but that they could come back the following day. This limitation of five trials per day was imposed to prevent participants from just mindlessly clicking in a mindless way.

The program kept a running record of the total number of trials, the number of hits, the number of first (and/or single) trials, and the number of hits on first trials. The experiment was planned to stop at 10,000 trials. Overall success rate was preplanned to be tested with a conservative two-tailed test and to perform a one-tailed chi-square to determine the difference in performance between first trials and subsequent trials. Trials were accumulated between June 1999 and May 2000.

**Results**

Altogether there were 5,482 hits in 10,000 trials ($\chi^2 = 1.46, \pi = 0.507, p = .14$, two-tailed), a hit rate of 34.02%, which is in the right direction but not significant. Contrary to prediction, the first trials ($N = 2,199$) performed somewhat worse than subsequent trials, $\chi^2(2, N=10000) = 0.17$ (wrong direction). The first trials obtained a hit rate of 33.65%. Detailed results are provided in Table 3.

**Table 3**

<table>
<thead>
<tr>
<th></th>
<th>First Trials</th>
<th>Subsequent Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>740</td>
<td>2662</td>
</tr>
<tr>
<td>Misses</td>
<td>1459</td>
<td>5139</td>
</tr>
<tr>
<td>Total</td>
<td>2199</td>
<td>7801</td>
</tr>
<tr>
<td>% hit rate (MCE 33.33%)</td>
<td>33.16%</td>
<td>34.12%</td>
</tr>
</tbody>
</table>

Contrary to prediction, the first trials performed somewhat worse than subsequent trials.

**Conclusion**

Although the overall results were not significant, a post hoc power analysis reveals that the power of the present experiment was only .26, assuming an effect size of $\pi = .1$. Given the small effect size, 108,213 trials are needed for a 95% probability of obtaining significant results. There are plans to conduct a larger study in the near future.

However, the results of the first trials were disappointing. Given Steinkamp’s (1998) effect size $\pi = .16$ for first trials, a significant effect should have been found with only 414 trials at the 95% confidence level. It is possible that, like the questionnaire data, the effect size was again inflated because of sampling error.

Web experimentation is still in its infancy in parapsychology. More experiments are required before it can be seen whether, or in which way, the Web can be a useful way of acquiring data.

---

1. Although, as a referee remarked, it may be that a low anxiety, dissociated state is psi-conducive.

2. Using the formula $\chi^2/N$, the effect size is .01 for all 10,000 trials. I used this estimate as roughly equivalent to $\pi$ to calculate the power analyses on C power. Strictly speaking, however, this effect size formula is not exactly equivalent to $\pi$, which is why $\pi$, a more suitable effect size measure for data comparing performance against chance, is reported in the body of the article.

3. However, Bierman has argued that larger studies generally produce smaller effect sizes in psi research (see http://ai162.feringa.nl/~eobh/publications/1996/freiburg1996.pdf). This may be because smaller studies provide inflated estimates of effect size, or there may be other factors (e.g., freshness of the experiment to the experimenter and/or to the participants) that promote psi in smaller studies specifically.
REFERENCES


Department of Psychology
University of Edinburgh
7 George Square
Edinburgh EH8 9JZ
Scotland, UK