Journal of Scientific Exploration, Vol. 24, No. 3, pp. 461–478, 2010 0892-3310/10

RESEARCH

Psi in a Skeptic's Lab
A Successful Replication of Ertel's Ball Selection Test

Suitbert Ertel

Georg-Elias-Müller-Institut für Psychologie. Waldweg 26, 37073 Göttingen, Germany
sertel@uni-goettingen.de

Abstract—In the Ball Selection Test for assessing psi, ping pong balls are drawn blindly from an opaque bag one at a time with replacement. Each ball has an integer from 1–5 and red or green dots marked on it, thereby producing 10 distinct alternatives. On each trial, a participant jumbles the balls, and attempts to guess both the number and the dot color on the ball prior to pulling it out of the bag. Because the 10 ball types are equally represented in the bag, the probability of correctly guessing both the number and the dot color by chance is 10%. In the full protocol, participants first test themselves at home without supervision. Those who score significantly above chance are then retested in the laboratory under an experimenter's supervision. In an experiment by the author with participants of the Georg-Elias-Müller Institute (GEMI), 47 participants achieved a hit rate of 11.6% in the at-home phase of the study, $p = 10^{-14}$ by a one-tailed binomial test; the 9 selected participants retested in the laboratory achieved a hit rate of 17.3% ($p = 10^{-50}$). A replication of the laboratory procedure was conducted by two graduating students working under the guidance of a skeptical professor at the Anomalistic Psychology Research Unit (APRU) at Goldsmiths College, University of London. Their 40 unselected APRU participants achieved a hit rate of 10.75, which was very significant by a binomial test ($p = .002$) and $p = .0003$ by summed Z^2 values. The lower hit rate of the APRU participants compared with GEMI participants was significant ($p = .02$) and predicted. It is argued that this low-tech testing procedure is less monotonous and more psi-conducive than conventional multiple choice procedures for testing psi.

Keywords: ESP—Ball Selection Test—skeptics

Introduction

Lack of replications of psi test results is a hotly debated problem of parapsychological research. Professor Chris French, head of the Anomalistic Psychology Research Unit (APRU), Goldsmiths College, University of London, offered me an opportunity to give a lecture to his team on the Ball Selection Test, a novel multiple choice procedure designed for screening psi abilities (Ertel, 2005a,

2005b, 2005c).[1] The participant's task in this test requires manual actions that are less monotonous than actions commonly required for conventional multiple choice procedures. Less tiresome conditions are generally regarded as more psi-conducive. I claimed that the Ball Test results are generally more replicable than those of other multiple choice procedures. Moreover, this test allows for considerably more trials per time unit; increased trial numbers improve statistical significance levels, if psi effects are real. I left a specimen of my test material at APRU, i.e. an ordinary opaque sports bag containing 50 ping pong balls as well as instruction and record sheets, in case members or students of French's team might want to give this test a try.

Shortly thereafter, two students at APRU, Johanna Körting and Luke Hagstrom, used this material for experiments that they conducted for a Final Year Project in fulfillment of BSc Psychology requirements (Hagstrom, 2002, Körting, 2002). After receiving their BSc certificates, they kindly provided me with copies of their theses and the data that they had collected. I wanted to know whether the students had replicated the results of the Ball Selection Test that I had obtained with participants at the Georg-Elias-Müller Institut für Psychologie (GEMI), Göttingen University, Germany. Two independent replications using this test had previously been made, one under my supervision by a student at GEMI who collected Ball Selection Test data for her diploma—later I thought that participants influenced by French's lectures might conduct this test with more reservation than had been the case with this GEMI thesis (Masuhr, 2000). Another study had been performed under the sole responsibility of researchers at IGPP (Institut für Grenzgebiete der Psychologie und Psychohygiene), Freiburg (published in Ertel, 2007). Both studies yielded very significant psi effects, even though the effect size was smaller than those that I, as experimenter, had obtained myself. The differences of results between my own study and the two replications might partly be due to "experimenter effects" (as investigated by Rhine & Pratt, 1957, Honorton, Ramsey, & Caribbo, 1975, Palmer, 1993, 1997, Schlitz & LaBerge, 1994, Watt & Ramakers, 2003, Smith, 2003). Psi manifestations are known to be affected by the experimenters' attitudes and personality traits. Open-mindedness of experimenters toward claims of the paranormal is deemed a favorable condition for psi effects, whereas skeptical attitudes seem to obliterate them. With some concern, I suggested to Professor French, editor of *The Skeptic,* that he encourage research with my Ball Test in his Unit.[2] I expected that experimenters would be from his staff and that the APRU students would be used as participants. I also presumed that the "skeptical look" by APRU would be mild enough and the Ball Test apt enough to let psi become manifest even under those suboptimal environmental conditions.

The following account of the APRU study is based on Körting's and Hagstrom's theses.[3] An independent re-analysis will be conducted first, and the

results of the two analyses (mine and the students') will be compared.

Methods

The students followed the standard procedure of the Ball Selection Test, as introduced at GEMI, Göttingen University. They had obtained GEMI's standard instructions (see Appendix A and Procedure below). Yet, the APRU study differed from GEMI studies in certain respects. The GEMI standard procedure has two stages. Stage 1: Participants, after receiving test material and instruction, complete the test at home, without supervision. After returning the filled-out record sheets and analyzing results, participants with significant home test scores—a minority out of the total—are invited to complete, under supervision, additional tests at the Institute, using the same material and instruction (Stage 2). The students at APRU and their teacher, however, decided to forgo home tests of participants. All tests were conducted under their supervision with unselected participants.

Experimenters and Participants

The experimenters, Luke Hagstrom and Johanna Körting, were undergraduate psychology students, aged 26 and 23, respectively. They tested 20 participants each.

The tests were completed by 40 participants, 14 male and 26 female, their ages ranged from 19 to 56, mean age 28.6. The majority were undergraduate students at APRU, most of them fellow students of the experimenters. Older family members were also tested. The participation was voluntary.

Materials

A participant's test material consists of 50 table tennis balls in an opaque bag with a drawstring opening. The drawstring is adjusted to make the opening wide enough to fit a hand into the bag. On each ball, one of the numbers 1, 2, 3, 4, or 5 is written in black permanent marker. Each number is written on 10 balls. A number is repeated ten times around one ball's sphere so as to make the number immediately visible irrespective of the ball's position. In addition, either red or green dots are marked on each ball in the space between the numbers; 5 balls of each number have red dots and 5 have green dots. This is the standard material of test version II which was used by Körting. Hagstrom used "air-flow" balls instead; these were of the same size as table tennis balls but with small holes. Half of the balls were white and half orange; the colors of the balls replaced the two-color marks of the table tennis balls. The change of material was, according to Hagstrom, due to "availability and time constraints". No further explanation was given.

Procedure

A participant's ball selection task consists of 6 runs. For each run, 60 trials are made within 10–15 min, on average. The total of 360 trials, completed in 60–90 min, was generally distributed over 2 or 3 sessions on different days. Since the participants were acquaintances of the experimenters, runs took place under familiar surroundings at the College or in student homes. The experimenter tried to create a relaxed atmosphere.

The participants were approached by oral invitation and were informed about the purpose of the experiment (a "test of extrasensory perception"). When they arrived for the test, individually, they were told that the data would be treated confidentially and that they were free to withdraw from the experiment at any time. They confirmed having received this information by signing a paper which gave informed consent. Only then were they asked to read the two pages of instructions (Appendix A) and to ask questions if they had any.

For each trial, participants place their left hand in the bag, left-handed individuals used their right hand, so as to have their writing hand free to record results immediately. The participant turns the bag over once in order to jumble the balls and to randomize their positions. Next, participants may proceed in one of two ways. They may single out a ball and, before taking it out, they guess which number is written on it and which color the dots have. Or they may first guess which number and color they will draw and then single out a ball. Participants are told that they may change the sequence of guessing and ball picking, but changes are only allowed between runs and are recorded by the experimenter.[4] After having taken a ball out of the bag, the experimenter records the selected number and color on the record sheet (Appendix B). Finally, the ball is put back into the bag, and the next trial begins.

The instructions for at-home testing of Phase 1 (Appendix A) direct that participants themselves record their guesses and the drawn numbers and colors. Hagstrom followed this same procedure for the laboratory tests at APRU, remaining inactive except for monitoring the participants to ensure that the procedures were correctly followed and that no cheating occurred. In contrast, Körting filled out the record sheets herself.

APRU participants also guessed prior to testing how they thought they would do and rated how they thought they had done after the testing. They also filled out personality and attitude questionnaires. But only the actual psi performance data are discussed in this article.

According to the instructions (Appendix A) (translated into English from GEMI's instructions for home test runs), participants jot down the guesses and selected numbers and colors. Hagstrom's participants complied with this: *"The experimenter was inactive throughout, except for watching the participants to ensure procedures were carried out correctly, and that cheating did not occur"*

(p. 12). Körting as experimenter was in charge of filling out the record sheets. She recorded the guesses as well as the drawn numbers and colors.[5] Before reading the instructions, the participants had to rate, on a five-point scale (*"very likely"* to *"very unlikely"*), whether they thought that they "will correctly guess, above chance, the numbers and/or colors written on the balls". After test completion, they had to select an answer to the question whether they believed they "correctly guessed the numbers/colors, above chance expectation. . . . " (*"I completely agree"* to *"I completely disagree"*). They then also completed an extraversion scale and a paranormal belief scale. The students' analyses also included analyses of correlations between hit scores in the ball test and personality data; Hagstrom was in charge of looking at extraversion, Körting of looking at possible "sheep and goat" effects (influence by belief). The student experimenters' pre- and post-experimental judgments and questionnaire data which are not part of the standard procedure had not been made available and are not considered in the present re-analysis. Hagstrom's original trial-by-trial data had not been provided either. He merely listed, for each participant, totals of number hits, color hits, and double hits for 360 trials (6 runs times 60 trials) (see Table 1). A trial-by-trial re-analysis can therefore be conducted only for Körting's participants.[6]

Results

The analyses as done by Hagstrom and Körting differed from what I would call appropriate. An account of an independent data analysis is provided first. The students' results are provided in detail below, for comparison.

1. Steps of analysis. The main variable of the ball test, version 2, which was used, is the double hit count, the count of hitting numbers and colors of single balls. Table 2 displays the results of the GEMI at-home tests on unselected participants, the GEMI laboratory tests on the selected subsample of participants successful at home, and the results of the APRU laboratory replication with unselected participants. Row 01 gives counts of participants, row 02 counts of trials which, divided by 60, yield the count of runs per participant. The sample's total trials are given in row 03. An analysis, using uni-directional tests of psi effects, yields Z_{bin}, which is based on scores summed across participants (row 04). The error probability p for this Z value is given in row 08, and the corresponding effect size ES_1 (for formula, see Equation (1) in the Table 2 legend) in row 09.

Since for the present data the observed hit rates per run (row 06) are all larger than expected (row 05), a one-tailed significance test (row 08) is considered appropriate. Negative deviations from chance (psi missing) were not hypothesized.

However, a more powerful way of analyzing psi test data of the multiple

TABLE 1
Summary Data for Each APRU Participant in Rank Order

Rank	Hit Count				
	N	***C***	***NC***	Z_{Bin}	***p***
1	89	224	67	5.36	10^{-6}
2	87	189	51	2.55	.007
3	75	180	50	2.37	.01
4	87	187	48	2.02	.03
5	90	189	47	1.84	.04
6	83	176	47	1.84	.04
7	68	187	46	1.67	.05
8	70	199	45	1.49	—
9	82	181	43	1.14	—
10	83	207	42	0.97	—
11	61	190	42	0.97	—
12	78	199	41	0.79	—
13	65	179	41	0.79	—
14	82	186	40	0.61	—
15	77	168	40	0.61	—
16	74	179	40	0.61	—
17	79	190	40	0.61	—
18	80	187	40	0.61	—
19	67	189	40	0.61	—
20	73	196	39	0.45	—
21	74	192	38	0.26	—
22	78	171	38	0.26	—
23	73	172	37	0.09	—
24	64	175	37	0.09	—
25	68	185	37	0.09	—
26	78	188	37	0.09	—
27	79	180	36	0.00	—
28	77	171	36	0.00	—
29	79	187	35	0.00	—
30	83	180	34	−0.26	—
31	68	180	34	−0.26	—
32	61	171	33	−0.44	—
33	55	195	32	−0.44	—
34	68	158	31	−0.79	—
35	52	184	29	−1.14	—
36	64	183	28	−1.32	—
37	58	146	28	−1.32	—
38	59	146	27	−1.49	—
39	72	171	26	−1.67	—
40	58	175	26	−1.67	—

N, C, NC, counts of number, color, and double hits (number plus color hits for individual balls).
Z_{Bin}, binomial *Z* values. *p*, one-sided significance level. —, not significant.

TABLE 2
Double-Hit Results of APRU Data Compared with Results from GEMI Studies with Unselected and Selected Participants

Double Hits Expected 10%	No.	Variables	APRU Unselected under Supervision	GEMI I Unselected without Supervision	GEMI II Selected under Supervision
Database essentials	01	$N_{Participants}$	40	47	9
	02	$Trials_{participant}$	360	480	480
	03	$Trials_{total}$	14,400	22,560	4,320
	04	Hit_{total}	1,548	2,620	748
	05	$Expected_{run}$	6.00	6.00	6.00
	06	$Observed_{run}$	6.45	6.97	10.39
Summed hits analysis	07	Z_{bin}	2.99	8.07	16.00
	08	p	**0.002	$^{**}10^{-14}$	$^{**}<10^{-50}$
	09	ES_1	0.025	0.054	0.243
Indicator bi-directionality	10	*Kurtosis*	5.95	5.93	2.10
	11	Z'	2.88	2.89	−0.27
	12	p	**0.002	**0.002	n. s.
Summed Z^2 analysis	13	Chi^2 $(df = N)$	76.3	288.0	279.4
	14	p	**0.0003	$^{**}<10^{-50}$	$^{**}<10^{-50}$
	15	Z''	3.43	>16.00	>16.00
	16	ES_2	0.029	0.067	>0.243

At GEMI, student assistants were in charge of experimenter control for seven participants, the author for two participants.
n.s., not significant.
**, very significant.

(1) $ES_1 = \dfrac{Z_{bin}}{\sqrt{Trials_{total}}}$ (2) $ES_2 = \dfrac{Z''}{\sqrt{Trials_{total}}}$ (3) $Chi^2 = \sum_{n=1}^{N} Z^2$

Z_{kurt} : Z of Kurtosis

Z_{Chi} : adapted from p of Chi^2 (D'Agostino, Belanger, & D'Agostino, 1990)

choice type has been provided by, among others, Timm (1983:222). Individual Z^2 values are used which, when summed across participants, yield Chi^2 values ($df = N$, number of participants, row 13). In contrast to using the summed hits procedure (rows 07–09), by using summed Z^2, participants with large psi-missing values contribute equally and positively to the psi indicator (Chi^2), the direction of deviations from MCE (mean chance expectancy) being irrelevant. Chi^2 can be transformed into equivalent Z values for the sample (row 15) and an effect size ES_2 is obtainable (see Equation (2) in Table 2) which may be compared, e.g., with effect size ES_1 (see Equation (1) in Table 2). The main advantage of individual Z score summation, with Z values squared, is that decisions between one- and two-sidedness of p tests are obsolete.[7]

An indicator of the kurtosis (curvature) of hit score distributions was also obtained. A significant kurtosis deviation from the expected value of 3.0, as

shown in rows 11 and 12 for APRU and GEMI I data, indicates a flat distribution, i.e. a non-normal spreading of hit scores in psi-hitting and psi-missing directions. A significant positive kurtosis would call for the Chi^2 analysis (summed Z^2). A Chi^2 analysis may not be needed, but is applicable no less, as a rule, with insignificant kurtosis deviations.

2. Summary of re-analyses of APRU data. Summed hit scores of APRU participants are very significantly larger than MCE ($p = .002$) by a one-sided test. The significance is more conspicuous ($p = .0003$) by summing Z^2 (Chi^2 test). The kurtosis of the data is flat (positive), which is noticeable by mere inspection of Table 1. Binomial Z values below zero (for participant ranks 30 to 40) increase rapidly in size, compared with the Z values above $Z = 0$. More indications of psi-missing tendencies of individuals in this sample are looked at below.

3. Comparing results of APRU and GEMI participants. Göttingen's GEMI I hit scores from 47 unselected participants, conducting the test under home conditions, are significantly larger than those from the 40 unselected APRU participants, who completed the test under supervision. Averages for one run are: 6.97 hits (GEMI) vs. 6.45 hits (APRU), respectively; expected are: 6 hits, see rows 05 and 06; the difference amounts to $Chi^2 = 6.55$, $df = 1$, $p = .01$.

GEMI II results of 9 participants, selected from the sample of 47 as good home test scorers and tested subsequently under control by one of two student assistants or by this author, obtained very large deviations from expectancy, compared with the total home test sample (averages 8.51 vs. 6.97 hits per run; expected: 6 hits). Not surprisingly, the hit score deviation of the selected $N = 9$ sample from the total of unselected $N = 47$ is very significant, $Chi^2 = 28.4$, $df = 1$, $p = 10^{-7}$.

The students' data analyses with commentaries. Both student experimenters subjected the scores of their $N = 40$ participants for number, color, and color-plus-number hits (= double hits), to one-sample t tests. The mean chance expectancy (MCE) of double hits of a participant was 36 (6 runs × 60 trials × 0.10 expected). However, t tests in this analysis are unsuitable because, to some extent, the size of deviations from expectancy (which is very important for assessing psi effects) is underrated while their variance (completely unimportant) bears upon the result. For example, with an MCE = 36 for each participant, a sample A of $N = 6$ participants obtaining, say, 37, 40, 43, 46, 49, 52 hits (total 266 hits) would obtain, with t test, a less significant p value for their successes ($t = 3.7$, $p = .007$) than an $N = 6$ sample B obtaining, say, 39, 40, 41, 42, 43, 44 hits (total 249 hits) ($t = 7.2$, $p = .0004$), while the actual deviation from chance probability of observed hit scores of sample A (266−216 = 50) is larger than that of sample B (249−216 = 33).

Even though a t test applied on the present data lacks power, both students obtained by two-sided tests significant or almost significant p's for double-hit

scores, the main psi indicator. Körting reports an almost significant $p = .052$; Hagstrom, for the same data, a significant $p = .038$. Hagstrom's two-sided p is correct, Körting's deviating value which should equal Hagstrom's is apparently due to a calculation error. However, the students should have applied a one-sided significance test since their "hypothesis one" has direction ("psi generates positive deviations from chance"). The correct one-sided error probability of observed APRU hit scores, obtained by a t test, is $p = .019$.

Hagstrom renders his t test insignificant by a Bonferroni correction since he claims that multiple indicators had been applied, i.e. aside from double hits (numbers plus colors) the significance also of number hits and color hits was obtained. But the Bonferroni correction is not admissible here since by instruction for participants the double-hit score is the ultimate success measure (see Instructions in Appendix A). A Bonferroni correction would have been admissible only for either number hits or color hits, if the double-hit score which the participants wanted to raise had not been significantly raised. Körting renders the (wrong) "almost significant" p insignificant as she concludes: *"None of the hit rates reached a level of statistical significance, implying that, from the overall results obtained in this study, one could not conclude that psi was operating for guessing of colors and numbers written on the balls"* (Körting, 2002:24). Her results of $p = .052$ which was not exactly significant should have been called "marginally significant".

Hagstrom, however, while taking a similar view of analysis, also did what should have been done in the first place. He calculated binomial Z values for participants individually and refers to the results briefly as follows: *"Although hypothesis 1 was not supported, the data offers support to the idea of psi existing with several participants reaching sig. hit levels ($p < .05$). This is particularly evident with the 'outlier'* [see participant rank #1]*. The probability of 224 or more color hits is* p = *.00000, of 89 or more number hits is* p = *.012056, and of 67 or more D [double] hits is* p = *.0000003"* (p. 16).

Hagstrom does not discuss the methodological difference between t test and binomial test, nor does he combine, which should have been done, the binomial Z values across participants. He concludes: *"Hypothesis one was offered no significant statistical support after data had been subjected to bonferroni adjustments"* (Hagstrom, 2002:17). Notwithstanding, his binomial p observations for individuals do not put it to rest: *"What proved to be the most interesting data came from participant no. 6 when binomial distribution was employed. 67 or more double hits had a probability of* p = *.0000003 (about 1 in 3 million). This although not in direct support of hypothesis one, does suggest that psi may exist."* The student does not simply verify a highly significant deviation from expectancy, he feels obliged to claim, at the same time, that this does not support hypothesis one ("psi exists"). Nevertheless, he continues: *"Psi may not be something we are all able to use or even possess. There may just be some gifted*

individuals.[8] *This idea is supported by data that Ertel (2000) has collected, with some of his participants achieving hit rates over 100% above MCE. When we look at most human abilities it is clear that people vary in how good they are, so why not psi*" (Hagstrom, 2002:18). Körting apparently did not look at individual differences of ball test scores; her final conclusion seems to be unambiguously consistent with possibly unsaid negative expectations: *"The results obtained by Ertel (2002) were not replicated"* (Hagstrom, 2002:30).

Discussion

The APRU students' way of analyzing their data (*t* tests) was inappropriate. By using appropriate statistical procedures, hypothesis one of the thesis writers ("psi effects exist") is confirmed. It should be added that the statement "psi effects exist" does not imply any commitment as to how many participants in the sample manifest psi effects individually and in which measure.

The APRU participants' hit scores, obtained under supervision, are significantly lower than results obtained, without supervision, from unselected GEMI participants. Apparently, in general, supervision is a psi-detrimental condition, probably due to increased emotional tension (for reviews, see White, 1976a, 1976b). In addition, differences between more optimistic (GEMI) vs. more skeptical social embeddings of the experiments (APRU) might have been effected.

The summed hit scores of 9 selected GEMI participants (GEMI II sample), tested under supervision after showing significant hit surplus under at-home conditions, were considerably larger than the hit scores of the 47 unselected GEMI participants who completed the test at home. On average, test participants, good at home, are still successful under supervision, but often with reduced effect size (Ertel, 2005b).

However, individual differences regarding hit scores under supervision are large. Under supervision, the double-hit scores for 7 out of the 9 participants of the GEMI II sample dropped more or less, compared with their home scores. But the scores of two participants, already high at home, increased noticeably under supervision—a surprising and as-yet-unexplained observation. At home the two high scorers obtained, with 480 trials, 80 and 89 double hits, respectively. Expected are 48 hits. Under supervision by the author they reached, again with 480 trials, 230 and 143 hits, respectively (230 is no typo!). These hit scores are unexpected and amazing.

An analysis of the Ball Selection Test data of version II (5 numbers and 2 colors as targets) considers double-hit scores as the crucial variable. Hit scores for numbers and colors alone may be of further interest, above all for looking at individual differences. One might want to know, e.g., whether participants differ regarding relative amounts of number vs. color hits. The correlations between

TABLE 3
Binomial *Z* Values for Number, Color, and Double Hits

			APRU Data (*N* = 40)			GEMI Data (*N* = 47)		
			Z_n	Z_c	Z_{nc}	Z_n	Z_c	Z_{nc}
	01	*a.M*	0.05	0.19	0.37	1.04	0.71	1.16
	02	*SD*	1.28	1.49	1.33	1.81	1.59	2.21
	03	Chi^2	65.7	90.6	76.3	204.4	140.2	288.6
r correlations	04	Z_n	1.00	0.40	0.66	1.00	0.30	0.83
	05	Z_{col}	0.40	1.00	0.63	0.30	1.00	0.56
	06	Z_{nc}	0.66	0.63	1.00	0.83	0.56	1.00
p of differences	07	Z_n	0.00	n.s.	n.s.	0.00	n.s.	n.s.
	08	Z_{col}	n.s.	0.00	n.s.	n.s.	0.00	n.s.
	09	Z_{nc}	n.s.	n.s.	0.00	n.s.	n.s.	0.00

Z_n, binomial Z of number hits only.

Z_c, binomial *Z* of color hits only.

Z_{nc}, binomial Z of double hits.

Rows 07–09, *p* values (two-tailed) are gained from two-sample paired *t* tests, $df = N - 1$.

number and color hits, regardless of occurring alone or in combination, are low (APRU: .40, GEMI: .30) (Table 3). Results for the two hit variables, numbers and colors separately, are given in Appendices C (Table 4) and D (Table 5) so as to enable researchers to make comparisons with future findings.

Skeptical observers of the ball test project tend to raise the objection that home test scores are worthless because they cannot be trusted. Their objection loses weight in view of the fact that success at home generally continues under supervision, even though, on average, at a somewhat lower level. Students serving honestly and cooperatively as participants for a scientist's research project hardly deserve to be looked upon with generalized suspicion. Bierman and Gerding (1991) should be mentioned as pioneers of studies with reduced experimenter control.

In view of the present results, more joint research of parapsychologists and their critics is called for. Professor French's cooperation was a rare exception. Die-hard skeptics want to make psi effects disappear altogether. They tend to do so by denaturalizing the test environment, i.e. by using dividing walls, blindfolds, and gloves; by letting participants pick each ball only once; by not giving feedback of hits and misses; by increasing numbers of skeptical onlookers and similarly psychologically inhibiting techniques (as observed by Ertel, 2007). Experimental successes at keeping the null-hypothesis unrejected would, however, hardly add much to our knowledge except by confirming that psi manifestations may be obstructed through inhibiting test conditions. The "does-psi-exist" question cannot be tested intelligently in this manner.

The test conditions as given in Appendix A should be revised only if experimental evidence would reveal that hit success above chance, as obtained under the Göttingen–London standard conditions, were due to sensory or memory leakage or some other non-psi factor. Several attempts to find evidence supporting a non-psi factor explanation for large hit rate deviations in the ball test had negative results (Ertel, 2004, 2005a, 2005b, 2005c). Success of further such attempts is deemed unlikely. Nevertheless, they should be continued.

Can fraud explain the results? If the answer is yes, then a considerable proportion of students in Göttingen must have been skilled conjurers. The student percentage with significant double-hit deviations from MCE was 32%, in London it was 15%. After subtracting a rate of chance expected, 5% from the successful Göttingen subsample of 27% (32%−5%) fraudulent students are obtained for Göttingen's campus, and 10% (15%−5%) for London's campus, if the observed hit rates were due to fraud. Such speculation disregards the reality of trust governing the majority of ordinary social interactions. Among students of science where curiosity usually gains the lead, fraud to raise hit scores in an experiment can only be expected to occur as an extremely rare exception.

Conclusion

Highly significant hit scores have been obtained by the author's Ball Selection Test applied under the supervision of an eminent British skeptic. The results, as analyzed by the student experimenters Hagstrom and Körting, who through inexperience were not fully conversant with the methods that should have been applied in the first place, nevertheless demonstrate that psi manifests itself even in a suboptimal environment.

This result should help to reduce a persistent reluctance to acknowledge the reality of the paranormal in skeptical circles. Moreover, the efficiency of this test for obtaining psi effects with a minimum of effort and expenditure might motivate researchers to try further replications. Parapsychological experiments are generally conducted, even today, without selecting psi-gifted participants. This might be the main reason why results in this discipline are often not replicable. Research is under way trying to further explore the validity of the ball test indicators. First correlations between ball test scores and other experimental psi manifestations have been obtained (convergent validity, Ertel, 2005a). The ball test might eventually serve as a valid tool for recruiting participants for studies in which a general psi ability is a desirable or even indispensable precondition.

Notes

[1] The original term "Ball Drawing Test" is ambiguous and may be replaced, as in this paper, with the "Ball Selection Test".

[2] From http://www.goldsmiths.ac.uk/departments/psychology/french.htm: "*The chal-*

lenge to those who adopt the working hypothesis that paranormal forces do not exist [Professor French's working hypothesis] is to provide plausible non-paranormal accounts, supported by strong empirical evidence wherever possible, of the ways in which psychological and physical factors might combine to give the impression that a paranormal event had occurred when, in fact, it had not." From the *The Skeptic* magazine's homepage: "*UK's only regular magazine to take a skeptical look at pseudoscience and claims of the paranormal."*

[3] I had preferred to publish this article with Professor French and the two (now) graduates as coauthors. Professor French declined coauthorship, but allowed use of the undergraduate experimenters' names, Körting and Hagstrom, who cannot be asked for consent as the APRU administration is unable to procure their addresses. Caroline Watt (2006) has shown that using undergraduate theses for re-analysis or surveys may be a profitable undertaking. Körting's and Hagstrom's theses provided on request.

[4] The participants' freedom of choosing between the two procedures is expected to increase their confidence. Their actual performance is deemed independent of procedural preferences, since sensory cues of the written numbers on the balls are lacking in both cases.

[5] This option is preferable since the experimenter's role, from the participants' perspective, is more meaningful and not apparently intrusive. It is actually also used in GEMI experiments in the lab, for second-stage experiments under supervision.

[6] Hagstrom provided summed individual hits for number, color, and double hits for 40 participants, in an Appendix of his thesis. Körting provided trial-by-trial data for her share of the sample, i.e. for 17 of her 20 participants only. The reason why the data of three participants were missing remains unknown.

[7] The reason is that Chi^2 tests are two-sided anyway. A one-sided p test for Chi^2 results would only be relevant if one predicted that the hit scores would cluster above chance close to MCE (yielding a negative kurtosis). But such prediction does not make sense.

[8] Hagstrom adds that the participant with extreme hit rates was his mother. It seems unreasonable to surmise that the young experimenter was deceived by his mother.

Acknowledgements

Stefan Schmidt helped, as *JSE*'s reviewer, to improve this paper considerably.

References

Bierman, D. J., & Gerding, J. L. F. (1991). Towards a reduction of experimenter control in the study of special subjects. *Proceedings of Presented Papers, The Parapsychological Association 35th Annual Convention.*

D'Agostino, R. B., Belanger, A., D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44,* 316–321.

Ertel, S. (2004). Are ESP test results mathematical artifacts? Brugger & Taylor's claims under scrutiny. *Journal of Consciousness Studies, 12,* 61–80.

Ertel, S. (2005a). The ball drawing test: Psi from untrodden ground. In: M. A. Thalbourne & L. Storm (Eds.), *Parapsychology in the Twentieth Century* (pp. 90–123). Jefferson NC: McFarland.

Ertel, S. (2005b). Psi test feats achieved alone at home: Do they disappear under lab control? *Australian Journal of Psychology, 10,* 149–164.

Ertel, S. (2005c). Are ESP test results stochastic artifacts? *Journal of Consciousness Studies, 12,* 61–80.

Ertel, S. (2007). Außersinnliche Wahrnehmung unter der Kontrolle organisierter Skeptiker. *Zeitschrift für Anomalistik, 7*(3), 236–269.

Hagstrom, L. (2002). *Psi and extraversion: An advancement of Ertel's Ball Drawing Test.* [Final year project BScPsychology thesis]. Supervisor Pr. Chris French. Goldsmiths College, University of London, Psychology Department.

Honorton, C., Ramsey, M., & Caribbo, C. (1975). Experimenter effects in extrasensory perception. *Journal of the American Society of Psychical Research, 69,* 135–139.

Körting, J. (2002). *The effects of paranormal belief on a new method for the investigation of psi: "The Ball Drawing Test"*. [BScPsychology Final year project thesis]. Supervisor Pr. Chris French. London: Goldsmiths College, University of London, Psychology Department.

Masuhr, B. (2000). *Persönlichkeit und die Neigung, sich dem Zufall zu entziehen*. [Diplomarbeit vom Georg-Elias-Müller-Institut für Psychologie]. Universität Göttingen.

Palmer, J. (1993). Confronting the experimenter effect. In: L. Coly & J. D. S. McMahon (Eds.), *Psi research methodology: A re-examination.* Proceedings of an international conference held in Chapel Hill, North Carolina, October 29–30, 1988 (pp. 44–64). New York: Parapsychology Foundation.

Palmer, J. (1997). The challenge of experimenter psi. *European Journal of Parapsychology, 13,* 110–125.

Rhine, J. B., & Pratt, J. G. (1957). *Parapsychology: Frontier Science of the Mind.* Springfield: Charles C. Thomas.

Schlitz, M., & LaBerge, S. (1994). Autonomic detection of remote observation: Two conceptual replications. *Proceedings of Presented Papers: The Parapsychologicval Association 37th Annual Convention* (pp. 352–360).

Smith, M. D. (2003). The psychology of the "psi-conductive" experimenter: Personality, attitudes towards psi, and personal psi experience. *Journal of Parapsychology*, *67,* 117–128.

Timm, U. (1983). Statistische Selektionsfehler in der Parapsychologie und in anderen empirischen Wissenschaften. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie, 2,* 195–229.

Watt, C. (2006). Research assistants or budding scientists? A review of 96 undergraduate student projects at the Koestler Parapsychology Unit. *Journal of Parapsychology, 70*(2), 335–356.

Watt, C., & Ramakers, P. (2003). Experimenter effects with a remote facilitation of attention focusing task: A study with multiple believer and disbeliever experimenters. *Journal of Parapsychology, 67,* 99–116.

White, R. A. (1976a). The influence of persons other than the experimenter on the subject's scores in psi experiments. *Journal of the American Society for Psychical Research, 70*(2), 133–166.

White, R. A. (1976b). The limits of experimenter influence on psi test results: Can any be set? *Journal of the American Society for Psychical Research*, *70*(4), 333–369.

Appendix A

Instructions for the Ball Selection Test

(Home Test Instructions)

The Ball Selection Test has been developed to test participants' paramental abilities. Paramental abilities are sometimes referred to with terms such as *psi sensitivity* or *sixth sense.* They may manifest themselves as sensory or motor accomplishments that cannot be explained by psychophysical mechanisms. In theory, all people might possess psi abilities. Their expression, however, often seems to be inhibited subconsciously. Inhibited psi effects remain untraceable, and this test cannot reveal conspicuous successes in this task despite general inhibition of pertinent subconscious dispositions.

The test material consists of 50 table tennis balls in an opaque bag. On each ball, one of the numbers 1, 2, 3, 4, or 5 is written. Ten balls carry number 1, 10 balls number 2, etc. Aside from the numbers, each ball carries either red or green dots. Among 10 balls with number 1, five balls have green dots, and five balls have red dots. The same

distribution of colored dots applies to the numbers 2, 3, 4, and 5.

You are provided with a record sheet [Appendix B] on which you will, please, note the results of this test. One sheet gives space for two runs.

For each run, when you begin, put down on the record sheet the actual date and clock time. Blanks for name and telephone number (and email address, if available) should also be filled out on the first record sheet.

Start off with a ball selection trial by putting your left arm into the bag (or your right arm if you are left-handed). The bag's opening should be adjusted so that you can move your arm in and out without much friction.

After putting your arm into the bag, you should jumble the balls so that their position becomes random. Turning the bag over on the table once, like a pancake on a pan, is an efficient jumbling technique. This should be done for each trial.

Next, single out a ball and draw it out of the bag. Before doing so, however, you guess and jot down on the record sheet the number and color that you may think you will draw on the next trial (see first row: "number/color"). For example, "3r" for trial (a) means "I expect/hope to draw number 3 with red dots on the ball". Use "r" for red and "g" for green.

After selecting the ball, see which number and color you have drawn, and jot down the number and color in the row underneath ("number/color drawn"). You might fail for both, number and color. You might hit the color but fail the number, or vice versa. Such partial hits are good, but of course hitting both number and color is the best result. Put an "H" in the third row for such double hits.

It is up to you how you guess the numbers and colors and how you select the balls. It might be helpful to close your eyes and visualize balls and numbers with colored dots. But no rule needs to be considered. You may open your mind for intuitions or concentrate mentally on your goal, you may grasp the balls in a straightforward way or surf over the balls trying to "feel" the numbers. You may try one ball first and then exchange it with another ball later, if you want, as long as you keep your hand in the bag. Peeping into the bag is not allowed. Raising your head as if you would look up to the ceiling, perhaps even with your eyes closed, would be the best way to avoid taking advantage, subconsciously, of visual cues.

Some people prefer first to grasp the ball and then make their guess. If you prefer this way of guessing, please make a note on the record sheet and maintain this procedure at least for the rest of that run.

After recording the number and color of the ball just drawn, please throw it back into the bag. It should bounce down, so do not leave it in your hand when you put it back for a next trial.

Sixty trials make one run. One run is divided on the record sheet into four rows of 15 trials. At the end of each row, please write down the number of double hits in that row. For data analysis, all hits will be considered, not only double hits.

While taking this test, you might have interesting experiences. Put down on the back of your sheets any observations that you think might be interesting to the experimenter.

If you would like to do some dry trials before filling out the record sheets, you may do four such trials without recording.

One run takes 10–15 min on average, first runs perhaps a little more. Your personal speed might deviate from the average; there is no time pressure.

Hit rates not only vary among participants, they also vary across sessions of individual participants. Considerable changes of hit rates of individuals across sessions cannot fully be explained, but participants tend to believe that low scores are due to bad mood, tenseness, overdrawn expectations, tiredness, and boredom. Please make a note on your record sheet if you suffer from any of the above conditions.

As soon as the analyses of your data is submitted, you will receive written feedback, if you wish. Thank you very much for your participation.

Appendix B

Record Sheet for the BALL SELECTION TEST

On the 1st sheet fill in your Name ________________ ***Tel:*** ________________

For each run jot down its number and date and time Email: ________________

Number of run:______ **Date:** ____________ **Time:** __________

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
Number / color guessed																
Number / color drawn:																Score
Mark double hits																
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
Number / color guessed																
Number / color drawn:																Score
Mark double hits																
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
Number / color guessed																
Number / color drawn:																Score
Mark double hits																
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	
Number / color guessed																
Number / color drawn:																Score
Mark double hits																

Appendix C

TABLE 4
Number Hit Results of APRU Participants Compared with Results from GEMI Studies

Number Hits Expected 20%	Number	Variable	APRU Unselected under Supervision	GEMI I Unselected without Supervision	GEMI II Selected under Supervision
Database	01	$N_{Participants}$	40	47	9
	02	$Trials_{participant}$	360	480	480
	03	$Trials_{total}$	14,400	22,560	4,320
	04	Hit_{total}	2,918	4,942	1,190
	05	$Expected_{run}$	12.00	12.00	12.00
	06	$Observed_{run}$	12.16	13.14	16.52
Summed hits analysis	07	Z_{bin}	0.78	7.15	12.38
	08	p	n.s.	$^{**}10^{-12}$	$^{**}10^{-35}$
	09	ES_1	0.007	0.048	0.188
Indicator bi-directionality	10	*Kurtosis*	2.22	4.45	1.62
	11	Z'	−1.16	2.01	−1.23
	12	p	n.s.	*.02	n. s.
Summed Z^2 analysis	13	$Chi^2\ (df = N)$	65.71	187.2	216.11
	14	p	**0.006	$^{**}10^{-13}$	$^{**}<10^{-35}$
	15	Z''	2.49	7.40	>12.38
	16	ES_2	0.021	0.049	>0.188

At GEMI, student assistants were in charge of experimenter control for seven participants, the author for two participants.
n.s., not significant.
**, very significant.

(1) $ES_1 = \dfrac{Z_{bin}}{\sqrt{Trials_{total}}}$ (2) $ES_2 = \dfrac{Z''}{\sqrt{Trials_{total}}}$ (3) $Chi^2 = \sum_{n=1}^{N} Z^2$

Z_{kurt} : Z of Kurtosis

Z_{Chi} : adapted from p of Chi^2 (D'Agostino, Belanger, & D'Agostino, 1990)

Appendix D

TABLE 5
Color Hit Results of APRU Participants Compared with Results from GEMI Studies

Color Hits Expected 50%	Number	Variable	APRU Unselected under Supervision	GEMI I Unselected without Supervision	GEMI II Selected under Supervision
Database	01	$N_{Participants}$	40	47	9
	02	$Trials_{participant}$	360	480	480
	03	$Trials_{total}$	14,400	22,560	4,320
	04	Hit_{total}	7,292	11,647	2,342
	05	$Expected_{run}$	30.00	30.0	30.00
	06	$Observed_{run}$	30.38	30.97	32.52
Summed hits analysis	07	Z_{bin}	1.53	4.88	5.52
	08	p	n.s.	$^{**}10^{-6}$	$^{**}10^{-7}$
	09	ES_1	0.013	0.032	0.084
Indicator bi-directionality	10	*Kurtosis*	4.76	3.62	3.38
	11	Z'	2.17	1.26	1.31
	12	p	0.01	n.s.	n. s.
Summed Z^2 analysis	13	$Chi^2\ (df = N)$	90.58	138.4	101.8
	14	p	$^{**}10^{-5}$	$^{**}10^{-10}$	$^{**}<10^{-17}$
	15	Z''	4.3	6.4	8.5
	16	ES_2	0.036	0.043	0.129

At GEMI, student assistants were in charge of experimenter control for seven participants, the author for two participants.
n.s., not significant.
**, very significant.

(1) $ES_1 = \dfrac{Z_{bin}}{\sqrt{Trials_{total}}}$ (2) $ES_2 = \dfrac{Z''}{\sqrt{Trials_{total}}}$ (3) $Chi^2 = \sum_{n=1}^{N} Z^2$

Z_{kurt} : Z of Kurtosis

Z_{Chi} : adapted from p of Chi^2 (D'Agostino, Belanger, & D'Agostino, 1990)