The Journal of Parapsychology, Vol. 61, December 1997

META-ANALYSIS OF FREE-RESPONSE ESP STUDIES WITHOUT ALTERED STATES OF CONSCIOUSNESS

BY JULIE MILTON

ABSTRACT: Seventy-eight free-response ESP studies in which participants were not in an altered state of consciousness were meta-analyzed. There was a highly significant cumulative effect (Stouffer $z = 5.72$, $p < 5.4 \times 10^{-9}$, one-tailed). The mean effect size ($z/N^{1/2}$) was 0.16 (*SD* = 0.29). Crude selective reporting appears to be an implausible counterexplanation of the overall cumulation and there were no strong positive indications that methodological artifacts played a role in producing above-chance results. However, the failure to report that the outcome measure was preplanned in 96% of the studies and the frequent use of multiple measures raises the possibility of widespread post hoc data selection problems within the studies themselves. No data are available to resolve the question of whether this threat to the database's validity is merely potential. Because of this problem and the obstacles to obtaining statistical evidence of the effects of flaws in any meta-analysis, caution is recommended in drawing strong conclusions from this meta-analysis about the existence of a genuine anomaly. However, a number of possible moderator variables were identified that suggest directions for future research.

Some of the strongest claims for the reality of extrasensory perception (ESP) in recent years have come from "remote viewing" studies. The term *remote viewing* is somewhat elastic in meaning, but, in recent years, has tended to be used to refer to free-response studies in which participants are not in an altered state of consciousness (ASC) and in which geographical locations or objects are the ESP targets. A number of unusually successful studies of this kind were reported early in the history of remote viewing by Targ and Puthoff (1977) and most recently by a research team funded by the US government to assess the usefulness of ESP for intelligence gathering (May et al., 1989). Both sets of studies have attracted the interest of other researchers seeking to replicate their findings, and both have been the center of controversy over whether their methodological quality is sufficient to justify claims that they constitute strong evidence for ESP (Marks & Kammann, 1980; May, 1996; Mumford et al., 1995; Wiseman & Milton, 1997).

Part of the material presented in this paper was reported at the Annual Convention of the Parapsychological Association, August 1993, at Toronto, Ontario, Canada.

I am grateful to Bob Morris and John Palmer for helpful comments on an earlier version of this paper, and to Jessica Utts and the late Chuck Honorton for valuable statistical advice and suggestions for variable coding.

Remote viewing studies are just a subset of a larger group of ESP studies in which free-response methods are used without the participants being in an altered state of consciousness. The claims made for remote viewing are alone sufficient reason to be interested in the larger database to which they belong, but an examination of free-response, nonASC studies also offers the opportunity to obtain some indication of whether the apparent success of ganzfeld ESP studies (Bem & Honorton, 1994) is due to the use of the ganzfeld technique or, as some have suggested, to the use of free-response, rather than forced-choice methods. Only two studies have compared a ganzfeld condition with a nonASC control condition (Murre et al., 1988; Palmer et al., 1980) and, in both studies, scoring in the two conditions did not differ significantly from each other. However, scoring in either condition also did not differ from mean chance expectation, so the failure to detect any difference might have merely been due to the absence of an overall effect.

The present meta-analysis of nonASC, free-response ESP studies was undertaken to address the following general questions: (a) Do the studies show statistically significant evidence of above-chance scoring? (b) What is the average effect size of the studies and how does it compare to that of the ganzfeld studies? (c) Is effect size related to the number of flaws within each study and/or to the presence of any particular flaw? Could any such significant relationship account for any overall significant deviation from chance in the data? (d) Do any participant characteristics or procedural variables moderate effect size?

The precise statistical formulation of these questions and a number of planned hypotheses relating to them will be detailed further on.

The Database

Study Designs

In free-response ESP studies, the receiver knows only what type of item the target might be (e.g., any picture, any object). The receiver attempts to describe the target by detailing the thoughts and images (mentation) that occur to him or her during the response period. A record of the mentation is compared with each item in a target set composed of the target and a number of decoys. Either the receiver or an independent judge, both blind to the target's identity, compares the mentation record to the target set items and ranks or rates them according to how well they match. Appropriate statistical tests are then applied to determine whether the target matches the mentation record better than the decoys to an extent beyond chance expectation.

Target sets can be generated in one of two ways. In an open-deck study, a large pool of potential targets is subdivided into fixed sets. The set to be used and the target within it are chosen randomly on each trial and a different set is used each time. In a closed-deck study, each trial's target is selected randomly from the large pool without replacement and at the end of the study the targets that were selected for each trial are combined to form a target set.

Study Retrieval

A number of bibliographic sources were searched for studies published between 1964 and January 1992, when the search was completed. The four main English-language parapsychology journals—*European Journal of Parapsychology, Journal of the American Society for Psychical Research, Journal of the Society for Psychical Research,* and *Journal of Parapsychology*—were surveyed for papers published in full. Abstracts of papers reported in *Research in Parapsychology* were also examined. Studies published in full in other journals were retrieved via *Parapsychology Abstracts International* and through a bibliography of remote-viewing studies (Hansen, Schlitz, & Tart, 1984).

Exclusion Criteria

Free-response ESP studies not involving ASCs were included in the meta-analysis. Prestated criteria excluded studies in which (a) participants were tested in groups; (b) the ESP task was "covert," i.e., disguised as another task such as a word-association test without participants being aware of its ESP aspect; (c) participants experienced an ASC such as dreaming or an ASC-induction procedure such as hypnosis or the ganzfeld, or some other procedure that the experimenter considered ASC-related; or (d) there were only four trials or fewer.

Meta-analysts sometimes prefer to keep severely flawed studies in a database and rely upon analyses contrasting flawed studies with unflawed studies to show up any effects that the flaws might have had on effect sizes. This approach has the potential advantages of increasing statistical power in the database both to detect a main effect and to determine whether flawed studies have higher effect sizes than unflawed studies. However, when only a very few studies in a database have a particular design flaw, the imbalance in size between the flawed group and unflawed group results in a surprisingly large reduction in statistical power for the contrast analysis. Rosenthal and Rosnow (1991) point out that when a *t* test is applied to a study with 100 trials, if that study has 95 trials in one group and 5 trials in the other, it effectively loses 81 trials compared with a study where the contrast groups are of 50 trials each. This means that the effects of rare flaws will have very little chance of being detected if the studies that contain them are included in a meta-analytic database of only moderate size. The inclusion of such studies could inflate the estimate of the main effect size and obscure the effects of important moderator variables on effect size.

When studies with unusual and potentially severe flaws were found, they were therefore not included in the present database. This principle led to the exclusion of a few studies in which the receiver was given an object belonging to a target person whom the receiver was asked to describe, and in which receivers were given deliberate sensory contact with nonidentical objects belonging to, or chosen by or for target persons, or were explicitly given information about target persons that might have related to the information about them that the receivers were attempting to guess. Two other studies with apparently severe flaws in design that did not apply to any other studies in the database were also excluded,[1] as was one in which the two participants later declared themselves to have been fraudulent in other experimentation.[2]

Also, several studies were excluded because they had reported only overall outcome measures based on inappropriate statistical procedures, which would have invalidated any effect-size measure derived from them.[3]

[1] The first study excluded was by Hearne (1986), who developed a novel design intended to address the challenge of testing spontaneous precognition in a nonlaboratory setting; unfortunately, the design essentially allows the receiver to select the target, nonrandomly and nonblind (Milton, 1988). The first Geller study by Targ and Puthoff (1974) was also excluded. This study used an unusual design in which the receiver was allowed to "pass" on any trial, before receiving target feedback, if he was not confident that the trial had been successful. However, Marks and Kammann (1980) provide evidence that the experimenters did not follow the protocol, and may have discarded the data from unsuccessful pass trials while keeping the data from successful ones. It should be noted that Marks and Kammann also criticized study IIIB of Puthoff and Targ (1976) for the possibility of a similar data selection problem, but later withdrew that criticism when provided with further information by Puthoff (see Marks, 1981, p. 200).

[2] The participants involved (Thalbourne & Shafer, 1983) signed a statement that they had not been fraudulent in this particular study, but since they had already shown themselves to be untrustworthy, it seems safer to exclude a study using known frauds.

[3] The most common flaw was to apply methods that inappropriately assumed that independent judges in closed-deck studies would judge each response transcript independently against each target set item (see Burdick & Kelly, 1977, pp. 113-114, for a discussion). This applied to studies by Allen et al. (1976); Bisaha and Dunne (1977); Honorton (1972); Krippner (1968); Musso and Granero (1973); Osis (1966); Puthoff and Targ (1976), Studies IIIC, D, and E; Rao and Feola (1973); Rauscher et al. (1976); and Reinsel and Wollman (1982). Puthoff and Targ (1976) report only an invalid analysis for Study IIIA, but a valid outcome measure for this study was reported in Puthoff and Targ (1975), and was used to calculate an effect size for the meta-analysis. The possibility of a "stacking effect" (Greville, 1944) ruled out Beloff and Mandleberg (1967), Moss and Gengerelli (1967) Group E, Smukler (1979), and Vallee (1988). Karnes et al. (1979), and Karnes et al. (1980) inappropriately used the number of independent judgings instead of the number of trials to calculate their outcome measure. Targ and Puthoff (1974) do not report what analysis was used to calculate their outcome measure in the second Geller study, and so

Characteristics of the Database

A study was predefined as a single experimental condition. Seventy-eight studies, reported in 55 papers by 35 different senior authors, were included in the database. There were 2,682 individual trials conducted overall (median of 16 trials per study, range 5 to 240), involving 1,158 receivers (median of 10 per study, range 1 to 74). The number of studies per senior author ranged from one to 11, with a median of two. Most studies (68%) used unselected participants.

Effect Size Measure

When a single outcome measure (direct hits, binary hits, etc.) was reported for a study, the standard normal deviate (*z* score) associated with the one-tailed probability of the outcome was obtained. Effect size was calculated by dividing the *z* score by the square root of the number of trials in the study. Studies merely reported as "nonsignificant" were assigned a *z* of zero.

It was anticipated that, as with early ganzfeld studies, many studies in the database would have used several outcome measures without stating which, if any, had been prespecified as the main measure and without correcting for multiple analysis (Hyman, 1985). Honorton (1985) addressed this problem in his ganzfeld meta-analysis by restricting his attention to direct hits as the outcome measure and discarding studies that did not report direct hits. However, nonASC free-response studies have been much more procedurally diverse than ganzfeld studies, and it seemed unlikely that many studies would have one outcome measure in common. It was therefore decided in advance that if a study reported more than one outcome measure, the mean value of the *z* scores associated with each measure would be used to calculate the study's effect size. This strategy can lead to an inflated estimate of effect size if the reported outcome measures are relatively successful ones, selected post hoc from among others. The potentially inflationary effect that using the mean effect size based on all the reported measures might have had will be examined below.

Flaw Criteria

Eighteen quality criteria were derived from criteria used in other parapsychological meta-analyses (Honorton, 1985; Honorton & Ferrari, 1989; Hyman, 1985; Hyman & Honorton, 1986; Radin & Ferrari, 1991) and suggested by methodological critiques (e.g., Akers, 1984; Hansel, 1966;

that study is excluded. Finally, there is a discrepancy between data provided in Table 2 (p. 153) of Tart and Smith (1968) for their second study, and the outcome measure derived from that data. The raw data no longer exist to be checked (Tart, personal communication, January 16, 1993), and so the study is excluded.

Kennedy, 1979a, b; Marks & Kammann, 1980; Schmeidler, 1977). Each study received one point for each safeguard and points were summed to give each study an overall quality rating. Telepathy, clairvoyance, and precognition studies were treated separately for the purposes of flaw analysis, because not all flaws apply to the three procedures.

The safeguard criteria that applied to all of the studies were as follows:

All Studies

Adequate randomization. Randomization of targets was considered adequate if the randomness source was a random number table, electronic random number generator, or tested mechanical device. Hand-shuffling, drawing lots, or casting dice were not considered adequate procedures. No quality credit was given if the randomization procedure was applied by participants rather than by experimenters, or if target selection involved an element of choice, as when someone is asked to choose as target the first "drawable" word on a page of a dictionary opened at random.

Random target position. Credit was given to studies in which the position of the target within the target set as presented to the judge was random. Credit was given to open-deck studies in which the target was randomly positioned in the target set; shuffling envelopes containing target set items or simply choosing where in the set to place items was not considered random. Credit was given to closed-deck studies in which targets and transcripts were in random order with respect to each other.

No handling cues. Credit was withheld from open-deck studies that did not use a duplicate of the target set for judging when the target was handled separately from the decoys.

Blind transcription. Credit was withheld when response transcripts were made by someone who was not specified as being blind to the target identity.

No past target cues. When the same receiver does more than one trial in a study and independent judging is used, safeguards are necessary to ensure that there is nothing in the response transcripts of later trials that might give clues to the identity of the target on earlier trials. Credit was given to studies that avoided this problem by *not* employing the following combinations of procedures, all of which involve independent judging:

1. Receivers were given trial-by-trial target feedback of the target's identity and a closed deck was used (because receivers might have avoided talking about the content of targets on previous trials, artificially leading to above-chance scoring).

2. Receivers were given trial-by-trial feedback.*and* an open-deck procedure was used *and* independent judges did not judge the transcripts in the same order as the trials *and* references to previous targets were not edited out (because receivers might have mentioned what the target had been on

a previous trial).

3. A closed deck was used *and* independent judges could have determined the order of the targets *and* references to previous trials and targets were not edited out (because receivers might have given clues to previous targets or the order of trials).

No cues to judges. Some studies use target material whose characteristics change according to time and circumstance, such as geographical locations or people. For example, when it is raining, the streets will be wet and a person may be wearing waterproof clothing. This could cause problems because the receiver's mentation may be affected by the same circumstances: for example, on a rainy day, the receiver might refer to rain. If the experimenters use a closed-deck procedure with trials on different days and give the independent judges target descriptions or photographs made on the day of each trial, the judge could match the day's target and response simply according to references to the weather. Credit was therefore withheld from studies using such changeable targets if the descriptions of the targets for independent judges were made on the day of the trial and if receivers' mentation transcripts were not edited to remove all references to changeable aspects of the target. Credit was also withheld if other cues were available to judges from the target. For example, if movie clips of various durations were used as targets and clip duration determined the length of the response period, then quality credit was withheld because judges could have matched targets and mentation transcripts on the basis of length.

Double-blind target and response recording. Credit was given to studies in which target and response were recorded double blind or were double-checked by another experimenter. Targets were assumed to have been recorded double-blind if they had been selected in advance of the experiment, and responses were assumed to have been recorded double-blind if independent judges were used, or if there were multiple trials in a session before the receiver-experimenter received target feedback.

Double-checked outcome. Credit was given when the outcome measure for each trial (i.e., whether it was a hit or a miss, etc.) was independently checked by a colleague.

Optional stopping prevented. Credit was given when the number of trials in the study were stated as preplanned, or when the receiver and receiver-experimenter in a single-receiver study stopped before getting target feedback, or when the study was a condition in an experiment with matched *N*s across conditions.

Preplanned outcome measure. Credit was given when the overall ESP outcome measure for the study was stated as having been planned before the experiment began.

Overall Cumulation

All analyses, including specification of dependent variables and the statistical tests to be applied, were planned in advance of data collection unless otherwise stated. Alpha was set at .05.

Effect Size

The mean effect size of the 78 studies was .16 ($SD = 0.29$) with a 95% confidence interval from .10 to .22. The median effect size was .06, the lower value reflecting the fact that the distribution of effect sizes was somewhat skewed so that it peaked around mean chance expectation, but had more upper-tailed values than lower-tailed ones. The frequency distribution is shown in the stem-and-leaf plot in Table 1, which can be read conveniently like a frequency histogram, but which also displays precise information about each effect size. Each effect size is represented by combining a leaf with its stem so that, for example, the digits to the right of the stem .2 indicate that effect sizes of .24, .25, .27, and .28 were obtained in the category of $.2 \leq r < .3$. The skewness in the distribution is likely to have been caused because 18 studies—almost a quarter of the database—were assigned an effect size of .00 because they were reported only as "nonsignificant." These nonsignificant studies tended to be much smaller than studies for which outcome data were given, with a mean of 13 and median of 10 trials for the nonsignificant studies, and a mean of 41 and median of 21 trials for the others. This makes it possible, though uncertain, that the nonsignificant studies had a distribution of effect sizes similar to the others but were too small to reach statistical significance. If this was so then the median effect size is likely to be an underestimate of the true effect size, and the mean is likely to be the better measure of central tendency for the database.

The mean of the studies' effect sizes weighted by the number of trials in each study was .09. Such an analysis would be generally expected to give a more accurate estimate of effect size, because larger studies yield more accurate point estimates of effect size and should therefore carry more weight, unless there is evidence that study size can affect effect size through boredom or practice effects, which was not the case in this database.

The mean value of z was 0.65 ($SD = 1.15$). The overall combined Stouffer z was 5.72, which is statistically significant ($p < 5.4 \times 10^{-9}$, one-tailed), according to prediction.

Homogeneity

Effect sizes in the total sample were significantly heterogeneous[4] (p =.0008, one-tailed). Removal of the three studies that contributed most to the heterogeneity (Braud, Davis, & Wood, 1979a, Experiment 6; Eisenberg, 1973, Beta Condition; Morris et al., 1978, Project 2) left effect sizes that were no longer heterogeneous at the .05 alpha level. The mean effect size of the remaining, homogeneous, sample was 0.17 (SD = .28) with a 95% confidence interval from .11 to .23. Removal of the three outliers thus had little impact on the mean effect size of the database. As predicted, the Stouffer z of the homogeneous sample remains statistically significant (z = 5.85, $p < 10^{-7}$, one-tailed).

Replication Across Investigators

A one-way ANOVA indicated no significant difference in effect size between studies with different principal authors, $F(34,43) = 1.57$, $p = .08$.

Filedrawer Analysis

Rosenthal (1991) offers a method for computing how many unreported or unretrieved null studies with an average z of 0 would be required to bring the observed level of statistical significance in a database down to a nonsignificant level (alpha of .05). For the present database of 78 studies, an additional 866 null studies would be required to cancel out the overall significance, that is, roughly 11 null studies for each study in the database.

Effect Size and Year of Publication

There was no significant correlation between effect size and year of publication: $r_p(76) = -.12$.

Comparison with the ganzfeld database

At the time that this meta-analysis was carried out, the ganzfeld studies meta-analyzed by Hyman (1985) and Honorton (1985) constituted the most obvious group to which to compare a nonASC database. Neither Honorton nor Hyman employed $z/N^{1/2}$ as their effect size measure but it can easily be calculated using direct-hit data provided by Honorton (1985, p. 84, Table A1) for the 28 ganzfeld studies in his meta-analysis. A post hoc test showed that the mean effect size for the ganzfeld database (mean = 0.26, SD = 0.38) was higher than that for the current database (mean = 0.16, SD = 0.29), but not significantly so, $t(104) = 1.49$.

[4] Following Radin and Nelson (1989), Hedges' (1987) homogeneity test for Cohen's d was applied to the data.

However, a recent meta-analysis of ganzfeld studies conducted outside Honorton's laboratory since the publication of Hyman and Honorton's (1986) methodological guidelines for ganzfeld studies (Milton &Wiseman, 1997a) indicates a near-zero effect size (mean effect size = .02, *SD* = 0.23) and an overall null cumulation (Stouffer $z = 0.87$, $p = .19$, one-tailed). Clearly, in contrast to this group of ganzfeld studies, mean effect size in the present database is superior but it is not readily apparent which group of ganzfeld studies, if either, should form the basis for comparison.

Study Quality

Reliability of Quality Coding

Ideally, all studies in a meta-analysis would have their features coded by a number of assessors, all blind to each study's outcome, in order to determine the accuracy of the coding. Unfortunately, no such research assistance was available to me when this meta-analysis was conducted and I was the sole, nonblind coder. However, after the meta-analysis was completed it was possible to have eight studies (approximately a tenth of the database) coded blind by another assessor in order to give an indication of coding reliability.[5] The studies were chosen from among the telepathy studies in the database so that all quality criteria and moderator variables would be

Table 2

Correlations between effect size and study quality within subgroups of studies

Study type[a]	r_p	*df*
Telepathy	−.044	42
selected participants[b]	.178	11
unselected participants	−.068	26
Clairvoyance	.059	25
selected participants	−.320	5
unselected participants	.249	18
Precognition	−.363	4

[a] One study's outcome measure included trials of different types, so only 77 studies were included in this analysis.

[b] Telepathy studies using mixed groups of selected and unselected participants are excluded from this analysis.

[5] I am grateful to Fiona Steinkamp for carrying out the independent coding.

Table 3
UNWEIGHTED AND QUALITY-WEIGHTED STOUFFER Z SCORES FOR TELEPATHY, CLAIRVOYANCE, AND PRECOGNITION STUDIES

		Stouffer *z*	
Type of Study	*N*	Unweighted	Weighted
Telepathy	44	5.46****	3.77***
Clairvoyance	27	0.99	0.88
Precognition	6	3.06**	2.08*

*$p < .02$; **$p < .002$; ***$p < .0001$; ****$p < 10^{-7}$: one-tailed

relevant and the studies were pseudorandomly selected so that each study was reported by a different principal author.

The percentages of studies on which we agreed and the phi coefficients of our agreement are presented in Table 4 (values of phi could not be calculated for some quality variables because one or the other coder coded all eight studies in the same way. The average percentage agreement per variable was 78 (*SD* = 16) and the average value of phi for the 11 variables for which it could be calculated, was .41 (*SD* = .37). It will be noted that some values of phi are quite low (two are slightly negative), indicating poor agreement. My own rechecking of my own coding indicates that I was self-consistent in applying the coding criteria, so either the blind coder needed some practice to apply the coding criteria consistently to these low-agreement variables, or she was applying them consistently but using a different interpretation of them or a different set of assumptions about the studies from my own. Whatever the reason, any conclusions involving these low-agreement variables should be cautious.

Global Quality Measures and Effect Size

The correlations between each study's effect size and overall quality rating are summarized in Table 2. Positive correlations indicate that large effect sizes were associated with high quality. Analyses were also conducted separately for selected and unselected participants in case the combined analyses might obscure a tendency for only selected participants to exploit methodological weaknesses in the studies. Too few precognition studies used only selected or unselected participants to allow meaningful separate analysis. No significant correlations were observed.

Quality-weighted *z* scores were calculated for each study and combined separately for telepathy, clairvoyance, and precognition studies. It was predicted that both telepathy and clairvoyance studies would have significant quality-weighted Stouffer *z* scores (the number of precognition studies was expected to be too small for them to yield significance). Telepathy and precognition studies each had a significant Stouffer-combined quality-weighted *z*, but, contrary to prediction, not the clairvoyance studies. The results are given in Table 3, with each group's unweighted Stouffer *z* for comparison.[6]

Specific Safeguards And Effect Size

Table 4 shows how often individual safeguards were reported and the differences in mean effect size between studies that did and studies that did not report them. When both variances and sample sizes of two groups being compared were very unequal, the nonparametric Mann-Whitney two sample rank tests replaced the planned *t* tests.

Only one safeguard's absence showed a significant and positive relationship to effect size: studies not reporting that transcription of mentation reports was blind had higher effect sizes than studies with this safeguard ($p = .004$, one-tailed). Blind transcription is a precaution against someone who knows the target identity for a trial being biased by their knowledge to mistranscribe, add, or omit material to the response transcript in such a way that might provide cues that would help the judge select the target. Applying the (somewhat conservative) Bonferroni correction to take into account the fact that this was one of 16 safeguard contrast analyses gives a nonsignificant adjusted *p* value of .064. A post hoc analysis of the 58 studies that incorporated the blind transcription safeguard showed that their Stouffer *z* remained highly statistically significant ($z = 3.83$, $p < 7 \times 10^{-5}$, one-tailed), with a filedrawer of 256 unpublished null studies necessary to reduce the effect to nonsignificance.

Only three studies specified a preplanned outcome measure. With such a small number of studies in the group, no meaningful comparison with the contrast group is possible, and the "flawless" studies are too small a

[6] In these analyses, each methodological safeguard was arbitrarily assigned a weight of one. However, it could be argued that some flaws are potentially more important than others, and should be given higher weights. Post hoc, ten researchers who were experienced free-response ESP experimenters or who had written methodological critiques of such studies were asked to rate each methodological safeguard in terms of how critical they felt it to be. Ratings were on a scale from one (*negligible importance*) to five (*critical importance*). However, there was very little variation in the importance that the experts assigned to each safeguard; 85% of the safeguards received a mean rating of 4.0 or over. Because these flaw ratings would have yielded weights very similar to the original unitary weights for the different safeguards, no analyses were conducted using the experts' ratings as weights.

Table 4
MEAN EFFECT SIZE AS A FUNCTION OF REPORTED SAFEGUARDS

		Mean effect size		Comparison		Intercoder agreement	
Safeguard	% studies reporting safeguard	Without safeguard	With safeguard	Test	z	%	(ϕ)
All study types (N = 78)							
Randomization adequate	49	.20	.12	t = 1.15	1.14	100	1.00
Target position random	79	.14	.17	t = –0.39	–0.39	50	.00
No handling cues	73	.08	.19	t = –1.54	–1.52	88	.77
Blind transcription	74	.34	.10	U = 1019.5	2.65	100	—
No past target cues	70	.27	.12	U = 1034.5	1.39	75	—
No cues to judges	91	.26	.15	t = 0.92	0.91	100	—
Double-blind recording	45	.18	.13	t = 0.79	0.79	63	.15
Doublechecked outcome	9	.16	.18	t = –0.17	–0.17	88	.75
Optional stopping prevented	59	.20	.13	t = 1.06	1.05	75	.49
Preplanned outcome measure	4	.15	.50	—	—	75	.49

Table 4, *continued*

Safeguard	% studies reporting safeguard	Mean effect size		Comparison		Intercoder agreement	
		Without safeguard	With safeguard	Test	z	%	(ϕ)
Telepathy and clairvoyance studies (N = 72)							
Sensory screening	71	.11	.16	$t =$ –0.73	–0.73	75	—
No intermediaries	82	.09	.16	$U =$ 428.5	–0.67	75	–.14
Experimenter blind	92	.09	.15	$U =$ 216.0	–0.05	88	—
Participants supervised	65	.08	.18	$U =$ 816.0	–1.15	88	.65
Confederate obstructed	19	.14	.15	$U =$ 2131.5	–0.20	75	.49
Materials secure	24	.13	.20	$t =$ –0.87	0.86	38	–.15
Telepathy studies (N = 45)							
Sender not initiator	84	.29	.17	$t =$ 1.06	1.05	75	—
No sender fraud	100	—	—	—	—	88	—

database from which to draw conclusions about whether nonASC, free-response studies in general show evidence for ESP. The implications of this situation will be addressed in the discussion section.

Overall Quality and Year of Publication

Study quality increased over time for telepathy and clairvoyance studies, as shown by correlations between quality rating and the year of publication, $r_p(42) = .70$, $p < 10^{-4}$ and $r_p(25) = .27$, *ns*, respectively, but decreased for the six precognition studies in the database, $r_p(4) = -.69$, *ns*; all tests two-tailed.

Overall Quality and Source of Publication

The main parapsychology journals do not impose severe space restraints on accounts of experiments, but fully 45% of the studies in the database were published as summaries, rather than as full papers, in *Research in Parapsychology*.[7] It is important to establish whether this might have led to an over-pessimistic assessment of database quality due to authors having little space in which to describe their studies' methodology. The same

Table 5
MEAN STUDY QUALITY RATINGS
(EXPRESSED AS PERCENTAGE OF MAXIMUM POSSIBLE FOR STUDY TYPE)
ACCORDING TO PUBLICATION SOURCE

	Study Type		
Publication Source	Telepathy	Clairvoyance	Precognition
Research in Parapsychology	62	56	50
Main parapsychology journals	62	60	43
Other journals	51	—	—

[7] Commenting on an earlier draft of this paper, John Palmer (personal communication, March 3, 1994) points out that in parapsychological meta-analyses a more accurate (and favorable) assessment of methodological quality would be expected by using the full versions of papers in *Proceedings of the Parapsychological Association Annual Convention* rather than the summaries of the papers contained in *Research in Parapsychology*. Surprisingly, in this database, cross-checking indicates that in only 2 out of 35 studies from *Research in Parapsychology* did the *Proceedings* give extra detail that would have increased a study's quality rating. However, this was largely due to the practice of not publishing research briefs, poster presentations, or versions of papers that differed from their summaries in the *Proceedings* in its early years and is no reason why meta-analysts should not follow Palmer's recommendation in future.

concern applies to the 6% of studies not published in the main parapsychology journals.

One-way ANOVAs were planned to test for differences in study quality according to whether a study was published in *Research in Parapsychology*, in one of the four main parapsychology journals, or in any other journal. No significant differences were found; surprisingly, mean quality ratings were very similar within study type for all publication sources. The data are summarized in Table 5.

Moderator Variables

The aim of this last group of analyses was to examine as many variables as possible considered important by reviewers and meta-analysts of parapsychological research (Carpenter, 1977; Delanoy, 1987; Honorton et al., 1990; Honorton & Ferrari, 1989; Milton, 1990; Morris, 1978; Palmer, 1978, 1986). Sixty-one comparisons and correlations were planned, but only in 31 cases were sample sizes large enough (at least five studies in any one comparison or correlational group) for meaningful analysis. The results are summarized in the Appendix. Two post hoc analyses are also reported, comparing selected versus unselected receivers and naive versus experienced receivers. These analyses had been planned to apply only to studies with one trial per receiver, but they could not be carried out because all such studies used naive, unselected receivers. In some cases, the assumptions underlying the use of the planned *t* tests were not met, and Mann-Whitney two sample rank tests were used instead.

Intercoder reliability data for each variable, based on the same sample of eight studies as were used for the quality coding check discussed earlier (see Footnote 4), are also presented in the Appendix. The percentages of studies on which the two coders agreed and the correlation coefficients of our agreement are presented in Table 4. In the case of these data, some correlations could not be calculated, not because a coder coded all eight studies the same (with the single exception of the variable relating to whether the experimenter told the receiver that success in the task was likely), but because the data were not binary and could not be meaningfully dummy-coded on a scale. Also, although some variables are continuous, the necessity of coding some studies as "unreported" for them ruled out the possibility of calculating correlation coefficients for those variables. In the binary quality coding data discussed earlier, a high percentage agreement alone does not indicate good intracoder reliability. If judges simply share a strong bias for assigning studies quality credit for a particular safeguard, then they will tend to agree on their ratings of many studies simply by chance. However, when more than two coding categories apply to a moderator variable for these eight studies, strong bias is less of a problem and percentage agreement alone is a reasonably good guide to intracoder reliability in the absence of a correlation coefficient. The mean percentage agreement for the 33 variables coded by both assessors was 79 (*SD* = 18), and the mean value of phi, for the 11 variables for which it could be calculated, was .73 (*SD* = .23).

Effect size did not vary significantly according to the type of test (telepathy, clairvoyance, or precognition), nor was it related to characteristics of receivers or senders. Honorton et al. (1990) reported that in the ganzfeld studies meta-analyzed by Honorton (1985), those telepathy studies in which receivers were free to bring friends to act as senders obtained significantly higher scoring than studies in which senders were assigned by the laboratory ($r = .40$, 23 *df*, $p = .023$). A correlation of similar size and direction ($r = .36$) was obtained for Honorton et al.'s 11 autoganzfeld studies. A prediction was therefore made for the present meta-analysis that effect sizes would be significantly higher when receivers and senders were friends than when they were strangers, but effect sizes were nonsignificantly lower.

One aspect of the random target selection procedure related significantly to effect size. Effect sizes were higher when random number tables were used than when "true" random number generators were employed, $r(31) = .35$, $p = .046$, two-tailed. Effect size did not relate significantly to the number of trials in the study or to the number of trials per receiver or per agent.

Two procedures for dealing with the receiver and sender during the trial yielded significant effects. Studies in which the receiver-experimenter (while still blind to the target identity) asked the receivers for fuller details of the mentation obtained higher effect sizes than other studies, $r(65) = .26$, $p = .036$, two-tailed. Effect sizes were significantly lower in studies in which senders were instructed to attempt to mentally convey the content of the target material to the receiver than when they were not so instructed, $r(30) = .35$, $p = .046$, two-tailed. It should be noted that this latter analysis was conducted only upon telepathy studies in which the author indicated that the person who knew the target identity was a sender, rather than just someone guarding the target or with incidental knowledge of its contents.

A one-way ANOVA indicated that the type of target material used was related to effect size and post hoc analysis suggests that this difference was essentially due to higher effect sizes in studies using objects and geographical sites as targets as compared to pictures. Objects and geographical locations were both individually significantly more associated with higher effect sizes than pictures, $r(66) = .41$, $p = .0006$, and $r(66) = .38$, $p = .0028$, respectively, with none of the other possible comparisons yielding significant outcomes.

Two significant outcomes were obtained in analyses related to the judging process. The number of items in the judging set correlated significantly and positively with effect size, $r(71) = .50$, $p = .000008$, two-tailed. Inspection of the data reveals that the correlation appears to be largely due to a wholesale upward shift in average effect size for sets containing more than eight items, rather than being due to a steady increase of effect size with set size. Higher effect sizes were obtained with independent judges than with receiver judges $r(73) = .40$, $p = .0006$, two-tailed).

Applying the Bonferroni correction to take into account that these statistically significant outcomes were obtained in the context of performing 33 moderator variable analyses, only the relationships between effect size and three variables—target type, the number of items in the judging set, and the use of independent versus receiver judges—would remain significant with an alpha of .05.

Variable Confounds

Leaving aside the question of multiple analysis, out of a total of 49 analyses of flaw-related and other potential moderator variables, 7 variables related significantly to effect size. If some of the variables are themselves intercorrelated, it may suggest that for some of them, their apparent relationship with effect size is an artifact of their being correlated with a variable that has a genuinely causative relationship to effect size.

Table 6 shows phi coefficients for the relationships between the seven variables. The value of each variable that was associated with the higher effect size was coded as "1," and the other in the pair as "0." Judging set size

Table 6

PHI COEFFICIENTS FOR VARIABLES THAT RELATED STATISTICALLY SIGNIFICANTLY TO EFFECT SIZE

Variable	1	2	3	4	5	6
1. Blind transcription flaw						
2. Randomization source	.20					
3. Sender instructions	.15	.19				
4. Extra mentation details	.52	.17	.12			
5. Target type	.28	.28	.05	.32		
6. Judging set size	.25	.08	.21	.19	.44	
7. Identity of judge	.58	.40	.30	.32	.56	.42

was dummy-coded as "0" for set sizes of eight or less, and "1" for set sizes greater than eight. It can be seen that all of the variables that were significantly related to effect size are positively correlated with each other, some quite strongly. The confounding of the blind transcription flaw with the eliciting of extra mentation details and the use of independent rather than receiver judges in particular (ϕ = .52 and .58, respectively) raises the possibility that the apparent effects of the latter two variables were not independent.

Discussion

Overall Cumulation

The evidence for an above-chance effect size in the 78 studies examined was highly statistically significant and did not appear to be due to a few exceptional studies or investigators. Selective reporting of successful studies seems unlikely to account for the overall effect. Stanford (1992) points out reasons why the usual calculations might lead to unrealistically large filedrawer estimates (see also Rosenthal, 1992), but it seems unlikely that there is an unpublished filedrawer of null studies even twice the size of the published database, let alone one 11 times its size as required by Rosenthal's (1991) method of estimation. It is relatively easy to publish studies with null results in parapsychology, and it has been Parapsychological Association policy since 1975 that journals should not discriminate against such studies (Parapsychological Association, 1975); indeed, in the current database, 77% of the studies reported non-significant results. Moreover, Hansen, Schlitz, and Tart (1984) and Blackmore (1980), attempting to locate unpublished remote viewing and ganzfeld studies, respectively, found unpublished databases roughly equal in size to the published databases. This suggests a similar situation in the present database, which includes the published studies from Hansen et al.'s search and is also likely to be similar to ganzfeld studies in terms of experimenters' time investment and consequent desire to publish. A filedrawer problem appears even more unlikely for the additional reason that neither survey found much difference in the proportions of published and unpublished reports reaching statistical significance with an alpha of .05 (54% and 44%, respectively, for the remote viewing studies, 58% and 37%, respectively, for the ganzfeld studies; mean effect sizes were reported for neither database).

Comparison with Ganzfeld Studies

The question of whether effect sizes in this database compare well to those obtained in ganzfeld studies is rather difficult to answer. The nonASC studies did not obtain a significantly lower mean effect size than the early ganzfeld studies but the lack of significance may have been due to the relatively small number of studies involved. However, the nonASC studies clearly outperform the near-zero mean effect size of the new ganzfeld studies. At this stage it is only possible to speculate on the reason for the decrease in effect size of the ganzfeld studies. It may be due to increased methodological stringency, or, perhaps, decreased use of psi-conducive procedures, or the studies may even have been the victims of a "meta-analytic demolition effect" as posited by Haraldsson and Houtkooper (1994). Until we know what factors are relevant, the question of whether the ganzfeld technique enhances effect sizes in free-response studies will probably be best addressed by meta-analyzing within-study comparisons of ganzfeld and nonganzfeld conditions with all other factors held constant.

Methodological Quality

Like any other database in any scientific discipline, this one is not perfect. In terms of reporting safeguard procedures, it is probably better than many databases in more mainstream areas of psychology. As in all meta-analyses, however, it is important to examine the possible role of methodological flaws in producing an apparent effect. There was no positive evidence that flaws played an important part in the overall cumulation. The number of procedural flaws in a study did not correlate significantly with effect size, and although studies that did not report blind transcription of receivers' mentation reports had significantly higher effect sizes than studies reporting this safeguard, the 58 studies that did report the safeguard still had an overall highly significant cumulated outcome well beyond what could be accounted for by selection from a null filedrawer of plausible size.

As with all meta-analyses, however, absence of evidence for relationships between effect size and study quality cannot be taken as evidence of absence. A thorough discussion of the ways in which procedural flaws might account for observed effects without revealing their actions through correlative analysis is beyond the scope of this paper, but a few examples will give an indication of some of the issues involved.

As a first example, a negative correlation between the presence of certain flaws might disguise their individual effects. Taking an extreme case to illustrate the point, if either flaw A or flaw B are always present in a study but never both together, and if they have similar effect sizes, then the individual effects of neither will be apparent and the total number of a study's flaws will not relate to effect size (Hyman, 1985). A second possibility is that a negative correlation between effect size and study quality could be obscured by an artifactual counter-correlation caused by experimenters who obtained null results feeling less obliged to report all their methodological safeguards. Third, the binary flaw coding (flaw absent vs. flaw present) used in this and most parapsychological meta-analyses is rather crude.

Studies that do not report whether they carried out a particular safeguard are grouped (usually) with studies that are clearly flawed. If studies without the safeguard have higher effect sizes and if many of these indeterminate studies did in fact use the safeguard, then their inclusion in the flawed group will tend to obscure the effect of the flaw. Fourth, if flaws are assigned to studies incorrectly, then any relationship between flaws and effect size could be obscured. Inter-coder reliability in this meta-analysis, assessed using a small sample of studies, was reasonably high for most variables for which it could be assessed but low and even slightly negative for a few variables. Some other problems with interpreting nonsignificant correlations between flaws and effect sizes are discussed in Stanford and Stein (1994).

Quite apart from these difficulties in interpreting the correlative evidence bearing on the effects of flaws, there is a problem in this database concerning the use of possible post hoc selection of outcome measures that could not be addressed correlatively. Fully 96% of studies in the database did not report having prespecified a main outcome measure. It is likely that some studies prespecified a measure without reporting that they did so, but it is very difficult to make an estimate of how many studies might have carried out this precaution. The frequency with which multiple measures were reported (58% of studies) with no acknowledgment of a possible multiple analysis or post hoc data selection problem makes it difficult to be confident that prespecification was as widely used as we might wish. The prevalence of reporting multiple measures also raises the possibility that some experimenters who only reported a single measure may have examined the outcome of several and reported only the most successful.

It is clear that post hoc selection of any of several possible free-response outcome measures could be inflationary, and a computer simulation by Hyman (1985) gives an indication of what the upper boundary of the effect of such selection would be on studies' outcomes. He simulated 2,000 thirty-trial, four-choice ganzfeld trials in which he randomly generated rank and rating data for each trial. He calculated for each trial the four common outcome measures used in ganzfeld studies (direct hits, binary hits, sums of ranks, and standardized ratings), and found that the probability of obtaining at least one significant outcome on any of these measures per study to be .152, over three times the usual alpha level of .05. Hyman did not report the average effect size associated with such a strategy, but clearly its effects are not negligible.

No clear strategy for assessing the potential consequences of this problem in this database presents itself. Although we have a worst-case limit on the problem (96% of studies at fault), this is likely to be an overestimate of the problem, quite possibly a severe one. On the other hand, we do not have data to provide a best-case limit to convince us that the problem is trivial. Hyman (1985) and Honorton (1985) attempted to address this problem in their ganzfeld database by restricting their attention to studies reporting direct hits. (Hyman also included studies using binary hits and binary coding.) However, if authors tended to report only their more successful outcome measures, then this strategy might not have overcome the difficulty. The studies reporting direct hits would have been those studies in which direct hits (or a closely correlated measure from which direct hits could be derived) had been the most successful measure. All that is possible to say of the current database is that the results should be treated with caution.

Moderator Variables

Type I error may account for at least some of the seven statistically significant relationships between effect size and other variables, although even conservative corrections for multiple analysis would not reduce the relationship between effect size and target type, the number of items in the judging set, and the judge's identity to nonsignificance. However, all of the variables that related significantly to effect size were themselves positively intercorrelated, making it possible that a single variable was responsible for most or all of these effects. If that variable was the "blind transcription" flaw variable, then the other variables are clearly not of interest, but this appears unlikely. The blind transcription variable had neither the most statistically significant relationship to effect size, nor did it account for the largest difference in effect size out of all the variables.

These statistical arguments may help to indicate which relationships may be genuine but it may also be useful from this point of view to examine which of the significant variables make sense in terms of our existing knowledge. Of the nonflaw-related variables, the relationships of effect size to the type of randomization source and to the sender's instructions appear the least likely to be meaningful. Both results are early casualties of correction for multiple analysis and there is no strong theoretical basis to predict their effects in the observed direction. The finding in this database that studies in which extra mentation details are elicited from receivers after the response period have higher effect sizes than other studies has support from within-study comparisons in the few studies that have made them (see Milton, 1990, for a review). In light of the finding that relatively large target sets can be made sense of, a larger target set would allow a relatively accurate receiver to obtain a higher effect size than a smaller target set. However, it could also be argued that at the typical levels of accuracy in ESP experiments, a large target set would make a near-miss lead to smaller effect sizes than a small target set. Given that there are arguments to be made in both directions for this variable, its apparent statistical significance should be treated cautiously. The same applies to the finding that the use of independent rather than receiver judges was associated with higher effect sizes in the database. Whether receiver or independent judges are superior has both arguments and empirical data from within-study comparisons in both directions (Milton, 1990). The apparent superiority of objects and geographical sites as targets can be interpreted meaningfully in terms of such targets being likely to make more discriminable judging sets than other targets, to motivate receivers more in the psi task, and to be more perceptually rich and so on. However, if pictures had turned out to be the most successful targets, then their superiority would also have been interpretable, in terms of their being more likely to be visually striking and emotional than other targets, for example.

However, it is notable that four elements of this cluster of variables—eliciting extra mentation details from receivers, using large judging sets, employing independent judges, and having objects and geographical sites as targets—represent the main features of a typical remote viewing experiment, although none of these procedures individually is unique to that paradigm. If one or more of these variables (including the absence of the blind mentation transcription safeguard) is not responsible for the higher effect sizes associated with this cluster, some other factor or factors typical of remote viewing studies may be a candidate. Experimenter effects are a possibility: Puthoff, Schlitz, and Russell Targ are three of the five most successful principal authors in terms of their studies' mean effect sizes, and they contributed 8 out of the 10 studies that employed all three of the procedures (large judging sets, objects or geographical sites as targets, and independent judges) that remain statistically significantly associated to effect size after correction for multiple analysis.

Other variables are also candidates for being the factor underlying the higher effect sizes in this variable cluster. Popular descriptions of remote viewing by the experimenters who originated the term (Targ & Puthoff, 1977) indicate that Targ and Puthoff's studies, at least, have some unusual features. The experimenters go to a great deal of trouble to make participants feel comfortable and valued in the laboratory setting and give them a high degree of favorable personal attention. Also, participants are not left alone to generate mentation but are interviewed during the response period by an experimenter who is blind to the identity of the target. This may give the experimenter the opportunity to divert the receiver from potentially unproductive styles of thinking such as simple free-association or attempts to reason what the target might be. Other experimenters using elements of remote-viewing procedures may also be using but not reporting some of these atypical procedures in sufficient numbers to account for the relatively high effect size in this subgroup of studies.

Recommendations for Future Research

Meta-analyses yielding spectacular probability cumulations, as this one does, have become prominent in parapsychology's attempts to demonstrate that its core phenomena are real. Certainly, meta-analyses have enabled parapsychologists to rule out with confidence the notion that the observed level of success can be accounted for by selective publishing of only successful studies. Even the most conservative of filedrawer estimates for most parapsychological meta-analyses are very implausible. Meta-analysis has also enabled parapsychologists to refute the assertion that there is a clear relationship between study quality and effect size visible, as it were, to the naked eye. As with this database, most meta-analyses of parapsychological databases show no positive evidence for such relationships in a way that can explain away the overall effects.

However, in databases in which potentially important safeguards are not reported and cannot be assumed to have been carried out, meta-analysis cannot itself offer a way to draw strong conclusions one way or another about whether the observed effects are genuine or artifactual. As with the more traditional literature review, the reader must make his or her own judgment about the importance of the various potential artifacts. In the present case, for example, a reader who considers that prespecifying a study's outcome measure is a matter of basic competence and likely to have been carried out even if not reported will come to a different conclusion from a reader who thinks otherwise. Some of this uncertainty could be reduced if future experimenters in this area describe their procedures in sufficient detail that it is clear that the studies meet quality criteria of the type presented here and elsewhere (Milton & Wiseman, 1997b).

If the effects detected in this meta-analysis represent a genuine communication anomaly rather than a methodological or statistical artifact, it will be important to optimize effect size in future studies. The fact that effect size has declined slightly over the years suggests that researchers have not been cumulatively applying process-oriented research to increase effect size in this type of study. The moderator variable analyses already discussed indicate some tentative pointers for a number of procedures. However, it is quite possible that variables that could not be examined in this meta-analysis are more important. Few studies reported details of procedures regarded as influential or even crucial by some commentators, such as instructions to participants, the nature of the experimenter's social interaction with participants, the characteristics of the target pool, and instructions to judges (Delanoy, 1989; Honorton, Ramsey, & Cabibbo, 1975; Milton, 1990; Watt, 1989; White, 1964). The variable-clustering problem demonstrated both within the database and indirectly in the attempted comparison with ganzfeld studies strongly suggests that any research investigating the effects of moderator variables should be conducted within studies, rather than between studies.

This meta-analysis has perhaps raised more questions than it has answered about whether there is convincing evidence for ESP from these nonASC free-response studies, and what variables, artifact-related or otherwise, might influence effect size. It may take a new generation of free-response studies to shed light on these issues, but effect-size estimates from the present database indicate that studies of the type examined here appear as promising as any in parapsychology.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Akers, C. (1984). Methodological criticisms of parapsychology. In S. Krippner (Ed.), *Advances in Parapsychological Research 4,* pp. 112–164. Jefferson, NC: McFarland.

Allen, S., Green, P., Rucker, K., Cohen, R., Goolsby, C., & Morris, R. L. (1976). A remote-viewing study using a modified version of the SRI procedure [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975,* (pp. 46–48). Metuchen, NJ: Scarecrow Press.

*Altom, K., & Braud, W. G. (1976). Clairvoyant and telepathic impressions of musical targets [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 171–174). Metuchen, NJ: Scarecrow Press.

*Bellis, J., & Morris, R. L. (1980). Openness, closedness and psi [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 98–99). Metuchen, NJ: Scarecrow Press.

Beloff, J., & Mandleberg, I. (1967). An attempted validation of the "waiting technique." *Journal of the Society for Psychical Research,* **44**, 82–88.

Bem, D. J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin,* **115**, 4–18.

*Bierman, D. J., Berendsen, J., Koenen, C., Kuipers, C., Louman, J., & Maissan, F. (1984). The effect of Ganzfeld stimulation and feedback in a clairvoyance task [Abstract]. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983,* p. 14. Metuchen, NJ: Scarecrow Press.

Bisaha, J. P., & Dunne, B. J. (1977). Precognitive remote viewing in the Chicago area: A replication of the Stanford experiment [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 84–86). Metuchen, NJ: Scarecrow Press.

Blackmore, S. J. (1980). The extent of selective reporting of ESP ganzfeld studies. *European Journal of Parapsychology,* **3**, 213–220.

*Braud, W. G. (1981). Psi performance and autonomic nervous system activity. *Journal of the American Society for Psychical Research,* **75**, 1–35.

*Braud, W. G., & Braud, L. W. (1975). The psi-conducive syndrome: Free-response GESP performance following evocation of "left-hemispheric" vs. "right-hemispheric" functioning [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1974* (pp. 17–20). Metuchen, NJ: Scarecrow Press.

*Braud, W., Davis, G., & Wood, R. (1979a). Experiments with Matthew Manning. *Journal of the Society for Psychical Research,* **50,** 199–223.

*Braud, W., Davis, G., & Wood, R. (1979b). "Psi on the tip of the tongue" revisited: A further investigation of the influence of an incubation period on free-response GESP [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 70–72). Metuchen, NJ: Scarecrow Press.

*Braud, W., & Masters, D. (1981). Psi performance and autonomic nervous system activity [Abstract]. In W. G. Roll & J. Beloff (Eds.), *Research in Parapsychology 1980* (pp. 124–128). Metuchen, NJ: Scarecrow Press.

*Braud, W. G., & Mellen, R. R. (1979). A preliminary investigation of clairvoyance during hypnotic age regression. *European Journal of Parapsychology,* **2,** 371–380.

*Braud, W. G., & Thorsrud, M. (1976). Psi on the tip of the tongue: A pilot study of the influence of an "incubation period" upon free response GESP performance [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 167–171). Metuchen, NJ: Scarecrow Press.

*Braud, W. G., Wood, R., & Braud, L. W. (1975). Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic Ganzfeld techniques: A replication and extension. *Journal of the American Society for Psychical Research,* **69,** 105–113.

Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of Parapsychology* (pp. 81–130). New York: Van Nostrand Reinhold.

Carpenter, J. C. (1977). Intrasubject and subject-agent effects in ESP experiments. In B. B. Wolman (Ed.), *Handbook of Parapsychology* (pp. 202–272). New York, NY: Van Nostrand Reinhold.

*Casler, L. (1982). "Active" versus "passive" GESP: A new approach. *Journal of the American Society for Psychical Research,* **76,** 167–176.

Delanoy, D. (1987). The reporting of methodology in ESP experiments. *Parapsychology Review,* **18,** 1–4.

Delanoy, D. (1989). Characteristics of successful free-response targets: Experimental findings and observations. In L. A. Henkel & R. E. Berger (Eds.), *Research in Parapsychology 1988* (pp. 92–95).

*Dunne, B. J., & Bisaha, J. P. (1979). Long distance precognitive remote viewing [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 68–70). Metuchen, NJ: Scarecrow Press.

*Eisenberg, H. (1973). Telepathic information transfer of emotional data [Abstract]. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology 1972* (pp. 134–136). Metuchen, NJ: Scarecrow Press.

*Gelade, G., & Harvie, R. (1975). Confidence ratings in an ESP task using affective stimuli. *Journal of the Society for Psychical Research,* **48,** 209–219.

*George, L. (1982). Enhancement of psi functioning through mental imagery training. *Journal of Parapsychology*, **46**, 111–125.

*Giesler, P. V. (1985). Parapsychological anthropology: II. A multi-method study of psi and psi-related processes in the Umbanda ritual trance consultation. *Journal of the American Society for Psychical Research*, **79**, 113–166.

*Giesler, P. V. (1986). GESP testing of shamanic cultists. *Journal of Parapsychology*, **50**, 123–153.

Greville, T. N. E. (1944). On multiple matching with one variable deck. *Annals of Mathematical Statistics*, **15**, 432–434.

Hansel, C. E. M. (1966). *ESP: A Scientific Evaluation.* NY: Scribner's.

Hansen, G. P., Schlitz, M. J., & Tart, C. T. (1984). Bibliography: Remote-viewing research 1973–1982. In R. Targ & K. Harary, *The Mind Race* (pp. 265–269). NY: Villard.

Haraldsson, E., & Houtkooper, J. M. (1994). Perceptual defensiveness, ESP, personality and belief: Meta-analysis, experimenter and decline effects. *Proceedings of Presented Papers: The Parapsychological Association 37th Annual Convention*, 161–174.

*Haraldsson, E., & Stevenson, I. (1974). An experiment with the Icelandic medium Hafsteinn Bjornsson. *Journal of the American Society for Psychical Research*, **68**, 192–202.

*Harding, S. E., & Thalbourne, M. A. (1981). Transcendental meditation, clairvoyant ability and psychological adjustment [Abstract]. In W. G. Roll & J. Beloff (Eds.), *Research in Parapsychology 1980* (pp. 71–73). Metuchen, NJ: Scarecrow Press.

Hearne, K. M. T. (1986). An analysis of premonitions, deposited over one year, from an apparently gifted subject. *Journal of the Society for Psychical Research*, **53**, 376–382.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, **42**, 443–455.

Honorton, C. (1972). Significant factors in hypnotically-induced clairvoyant dreams. *Journal of the American Society for Psychical Research*, **66**, 86–102.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology*, **49**, 51–91.

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., & Ferrari, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology*, **54**, 99–139.

Honorton, C., & Ferrari, D. C. (1989). Meta-analysis of forced-choice precognition experiments. *Journal of Parapsychology*, **53**, 281–308.

Honorton, C., Ramsey, M., & Cabibbo, C. (1975). Experimenter effects in extrasensory perception. *Journal of the American Society for Psychical Reseach*, **69**, 135–150.

Hyman, R. (1985). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology*, **49**, 3–49.

Hyman, R., & Honorton, C. (1986). A joint communique: The psi ganzfeld controversy. *Journal of Parapsychology*, **50**, 351–364.

*IRWIN, C. P. (1982). The role of memory in free-response ESP studies: Is target familiarity reflected in the scores? *Journal of the American Society for Psychical Research*, **76**, 1–22.

*KAPPERS, J. (1983). Screening for good ESP subjects with object-reading [Abstract]. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982* (pp. 150–151). Metuchen, NJ: Scarecrow Press.

KARNES, E. W., BALLOU, J., SUSMAN, E. P., & SWAROFF, P. (1979). Remote viewing: Failures to replicate with control comparisons. *Psychological Reports*, **45**, 963–973.

KARNES, E. W., SUSMAN, E. P., KLUSMAN, P., & TURCOTTE, L. (1980). Failures to replicate remote viewing using psychic subjects. *Zetetic Scholar*, **6**, 66–76.

KENNEDY, J. E. (1979a). Methodological problems in free-response ESP experiments. *Journal of the American Society for Psychical Research*, **73**, 1–15.

KENNEDY, J. E. (1979b). More on methodological issues in free-response psi experiments. *Journal of the American Society for Psychical Research*, **73**, 395–401.

*KESNER, J., & MORRIS, R. L. (1978). A precognition test using guided imagery [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1977* (pp. 48–52). Metuchen, NJ: Scarecrow Press.

KRIPPNER, S. (1968). Experimentally-induced telepathic effects in hypnosis and non-hypnosis groups. *Journal of the American Society for Psychical Research*, **62**, 387–398.

*MAHER, M. (1984). Correlated hemispheric assymetry in the sensory and ESP processing of continuous multiplex stimuli [Abstract]. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology 1983* (pp. 18–21). Metuchen, NJ: Scarecrow Press.

MARKS, D. (1981). On the review of *The Psychology of the Psychic:* A reply to Dr. Morris. *Journal of the American Society for Psychical Research*, **75**, 197–203.

MARKS, D., & KAMMANN, R. (1980). *The Psychology of the Psychic.* Buffalo, NY: Prometheus Books.

MAY, E. C. (1996). The American Institutes for Research Review of the Department of Defense's Star Gate Program: A commentary. *Journal of Parapsychology*, **60**, 3–23.

MAY, E. C., UTTS, J. M., TRASK, V. V., LUKE, W. L. W., FRIVOLD, T. J., & HUMPHREY, B. S. (1989). *Review of the psychoenergetic research conducted at SRI International (1973–1988). Final Report—Task 6. 0. 1, Project 1291.* Menlo Park, CA: SRI International.

*MCLENON, J., & HYMAN, R. (1987). A remote viewing experiment conducted by a skeptic and a believer. *Zetetic Scholar*, **12/13**, 21–33.

MILTON, J. (1988). Letter to the Editor. *Journal of the Society for Psychical Research*, **55**, 44–46.

MILTON, J. (1990). A survey of free-response judging practices. *Journal of the American Society for Psychical Research*, **84**, 189–225.

MILTON, J. & WISEMAN, R. (1997a). Ganzfeld at the crossroads: A meta-analysis of the new generation of studies. *The Parapsychological Association 40th Annual Convention: Proceedings of Presented Papers.* Hatfield, UK: University of Hertfordshire Press.

MILTON, J., & WISEMAN, R. (1997b). *Guidelines for extrasensory perception research.* Hatfield, UK: University of Hertfordshire Press.

*MOCKENHAUPT, S., ROBBLEE, P., NEVILLE, R. C., & MORRIS, R. L. (1977). Relaxation techniques, feedback, and GESP: A preliminary study [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1976* (pp. 50–52). Metuchen, NJ: Scarecrow Press.

MORRIS, R. L. (1978). A survey of methods and issues in ESP research. In S. Krippner (Ed.), *Advances in Parapsychological Research 2: Extrasensory Perception* (pp. 7–58). New York, NY: Plenum Press.

*MORRIS, R. L., & BAILEY, K. L. (1979). A preliminary exploration of some techniques reputed to improve free response ESP [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1978* (pp. 63–65). Metuchen, NJ: Scarecrow Press.

*MORRIS, R., ROBBLEE, P., NEVILLE, R., & BAILEY, K. (1978). Free-response ESP training with feedback to agent and receiver [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1977* (pp. 143–146). Metuchen, NJ: Scarecrow Press.

*MOSS, T. (1969). ESP effects in "artists" contrasted with "non-artists." *Journal of Parapsychology,* **33,** 57–69.

*MOSS, T., & GENGERELLI, J. A. (1967). Telepathy and emotional stimuli: A controlled experiment. *Journal of Abnormal Psychology,* **72,** 341–348.

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

MUMFORD, M. D., ROSE, A. M., & GOSLIN, D. A. (1995). *An evaluation of remote viewing: Research and applications.* Washington, DC: American Institutes for Research.

*MURRE, J. M. J., VAN DALEN, A. C., DIAS, L. R. B., & SCHOUTEN, S. A. (1988). A Ganzfeld psi experiment with a control condition. *Journal of Parapsychology,* **52,** 103–125.

MUSSO, J. R., & GRANERO, M. (1973). An ESP drawing experiment with a high-scoring subject. *Journal of Parapsychology,* **37,** 13–36.

OSIS, K. (1966). Linkage experiments with mediums. *Journal of the American Society for Psychical Research,* **60,** 91–124.

PALMER, J. (1978). Extrasensory perception: Research findings. In S. Krippner (Ed.), *Advances in Parapsychological Research 2: Extrasensory Perception* (pp. 59–243). New York, NY: Plenum Press.

PALMER, J. (1986). ESP research findings: The process approach. In H. L. Edge, R. L. Morris, J. Palmer, & J. H. Rush, *Foundations of Parapsychology: Exploring the Boundaries of Human Capability* (pp. 184–222). Boston, MA: Routledge & Kegan Paul.

*PALMER, J., WHITSON, T., & BOGART, D. N. (1980). Ganzfeld and remote viewing: A systematic comparison [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 169–171). Metuchen, NJ: Scarecrow Press.

PARAPSYCHOLOGICAL ASSOCIATION (1975). Motion concerning editorial and institutional policy. *Journal of Parapsychology,* **4,** 368.

*PRATT, J. G. (1966). New ESP tests with Mrs. Gloria Stewart. *Journal of the American Society for Psychical Research,* **60,** 321–339.

*Puthoff, H. E. (1985). ARV (associational remote viewing) applications [Abstract]. In R. A. White & J. Solfvin (Eds.), *Research in Parapsychology 1984* (pp. 121–122). Metuchen, NJ: Scarecrow Press.

*Puthoff, H. E., & Targ, R. (1975). Remote viewing of natural targets [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1974* (pp. 30–32). Metuchen, NJ: Scarecrow Press.

*Puthoff, H. E., & Targ, R. (1976). A perceptual channel for information transfer over kilometer distances: Historical perspective and recent research. *Proceedings of the Institute of Electrical and Electronic Engineers,* **64,** 329–354.

*Puthoff, H. E., Targ, R., & Tart, C. T. (1980). Resolution in remote-viewing studies: Mini-targets [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 120–122). Metuchen, NJ: Scarecrow Press.

Radin, D. I., & Ferrari, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration,* **5,** 61–73.

Radin, D. I., & Nelson, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics,* **19,** 1499–1514.

Rao, K. R., & Feola, J. (1973). Alpha rhythm and ESP in a free-response situation [Abstract]. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology 1972* (pp. 141–144). Metuchen, NJ: Scarecrow Press.

Rauscher, E. A., Weissmann, G., Sarpatti, J., & Sirag, S.-P. (1976). Remote perception of natural scenes, shielded against ordinary perception [Abstract]. In J. D. Morris, W. G. Roll, & R. L. Morris (Eds.), *Research in Parapsychology 1975* (pp. 41–45). Metuchen, NJ: Scarecrow Press.

Reinsel, R., & Wollman, M. (1982). A clairvoyance procedure using the semantic differential [Abstract]. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology 1981* (pp. 166–167). Metuchen, NJ: Scarecrow Press.

*Roll, W. G. (1971). Free verbal response and identi-kit tests with a medium. *Journal of the American Society for Psychical Research,* **65,** 185–191.

*Roll, W. G., Morris, R. L., Damgaard, J. A., Klein, J., & Roll, M. (1973). Free verbal response experiments with Lalsingh Harribance. *Journal of the American Society for Psychical Research,* **67,** 197–207.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research.* Newbury Park, CA: Sage.

Rosenthal, R. (1992). Effect size estimation, significance testing, and the file-drawer problem. *Journal of Parapsychology,* **56,** 57–58.

*Schlitz, M. J., & Braud, W. G. (1989). Free response psi performance with and without feedback: An attempted replication [Abstract]. In L. A. Henkel & R. E. Berger (Eds.), *Research in Parapsychology 1988* (pp. 53–58). Metuchen, NJ: Scarecrow Press.

*Schlitz, M., & Deacon, S. (1980). Remote viewing: A conceptual replication of Targ and Puthoff [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 124–126). Metuchen, NJ: Scarecrow Press.

*Schlitz, M., & Gruber, E. (1980). Transcontinental remote viewing. *Journal of Parapsychology,* **44,** 305–317.

*Schlitz, M., & Gruber, E. (1981). Transcontinental remote viewing: A rejudging. *Journal of Parapsychology,* **45,** 233–237.

*SCHLITZ, M. J., & HAIGHT, J. (1984). Remote viewing revisited: An intrasubject replication. *Journal of Parapsychology,* **48**, 39–49.

SCHMEIDLER, G. R. (1977). Methods for controlled research on ESP and PK. In B. B. Wolman (Ed.), *Handbook of Parapsychology* (pp. 131–159). New York: Van Nostrand Reinhold.

*SHAFER, M. (1982). Self-actualization, mystical experience, and clairvoyant ability: A second correlational test [Abstract]. In W. G. Roll, R. L. Morris, & R. A. White (Eds.), *Research in Parapsychology 1981* (pp. 153–155). Metuchen, NJ: Scarecrow Press.

SMUKLER, H. (1979). A remote-viewing experiment: California to Rhode Island. *Meta-Science Quarterly,* **1**, 25–32.

STANFORD, R. G. (1992). Experimental hypnosis-ESP literature: A review from the hypothesis-testing perspective. *Journal of Parapsychology,* **56**, 39–56.

*STANFORD, R. G., & MAYER, B. (1974). Relaxation as a psi-conducive state: A replication and exploration of parameters. *Journal of the American Society for Psychical Research,* **68**, 182–191.

*STANFORD, R. G., & PALMER, J. (1973). Meditation changed from medition—checkprior to the ESP task: An EEG study with an outstanding subject [Abstract]. In W. G. Roll, R. L. Morris, & J. D. Morris (Eds.), *Research in Parapsychology 1972* (pp. 34–36). Metuchen, NJ: Scarecrow Press.

*STANFORD, R. G., & PALMER, J. (1975). Free-response ESP performance and occipital alpha rhythms. *Journal of the American Society for Psychical Research,* **69**, 235–243.

STANFORD, R. G., & STEIN, A. G. (1994). A meta-analysis of ESP studies contrasting hypnosis and a comparison condition. *Journal of Parapsychology,* **58**, 235–269.

*TARG, E., & TARG, R. (1986). Accuracy of paranormal perception as a function of varying target probabilities. *Journal of Parapsychology,* **50**, 17–27.

*TARG, E., TARG, R., & LICHTARGE, O. (1985). Realtime clairvoyance: A study of remote-viewing without feedback. *Journal of the American Society for Psychical Research,* **79**, 493–500.

TARG, R., & PUTHOFF, H. E. (1974). Information transmission under conditions of sensory shielding. Nature, 252, 602–607.

TARG, R., & PUTHOFF, H. E. (1977). *Mind-Reach: Scientists Look at Psychic Ability.* New York, NY: Delacorte.

*TARG, R., PUTHOFF, H. E., HUMPHREY, B. S., & TART, C. T. (1980). Investigations of target acquisition [Abstract]. In W. G. Roll (Ed.), *Research in Parapsychology 1979* (pp. 122–124). Metuchen, NJ: Scarecrow Press.

*TART, C. T., & SMITH, J. (1968). Two token object studies with Peter Hurkos. *Journal of the American Society for Psychical Research,* **62**, 143–157.

THALBOURNE, M. A., & SHAFER, M. G. (1983). An ESP drawing experiment with two ostensible psychokinetes [Abstract]. In W. G. Roll, J. Beloff, & R. A. White (Eds.), *Research in Parapsychology 1982* (pp. 62–64). Metuchen, NJ: Scarecrow Press.

*TORNATORE, R. P. (1984). The use of fantasy in a children's ESP experiment [Abstract]. In R. A. White & R. S. Broughton (Eds.), *Research in Parapsychology*

1983 (pp. 102–103). Metuchen, NJ: Scarecrow Press.

VALLEE, J. (1988). Remote viewing and computer communications—an experiment. *Journal of Scientific Exploration*, **2**, 13–27.

*VENTURINO, M. (1978). An investigation of the relationship between EEF alpha activity and ESP performance. *Journal of the American Society for Psychical Research*, **72**, 141–152.

WATT (DOW), C. (1989). Characteristics of successful free-response targets: Theoretical considerations. In L. A. Henkel & R. E. Berger (Eds.), *Research in Parapsychology 1988* (pp. 95–99). Metuchen, NJ: Scarecrow Press.

WHITE, R. A. (1964). A comparison of old and new methods of response to targets in ESP experiments. *Journal of the American Society for Psychical Research*, **58**, 21–56.

*WIKLUND, N., & JACOBSON, N. O. (1976). A public experiment with precognition. *Journal of the American Society for Psychical Research*, **48**, 293–300.

WISEMAN, R., & MILTON, J. (1997). Experiment One of the SAIC remote viewing program: A critical re-evaluation. In R. Wiseman (Ed.), *The Parapsychological Association 40th Annual Convention: Proceedings of Presented Papers*. Hatfield, UK: University of Hertfordshire Press.

*WOOD, R., KIRK, J., & BRAUD, W. (1977). Free response GESP performance following Ganzfeld stimulation vs. induced relaxation, with verbalized vs. nonverbalized mentation: A failure to replicate. *European Journal of Parapsychology*, **1**, 80–93.

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

Appendix

Relationships between effect size and moderator variables, and intercoder agreement on assigning studies to each variable category

		Comparison		Intercoder agreement	
Variable	Mean effect size (*SD*)	Test	z	%	ϕ
Study type					
Precognition	.34 (.36)	$F(2,74)$ = 2.47	1.69	—	—
Telepathy	.18 (.29)				
Clairvoyance	.08 (.25)				
Receiver sampling characteristics					
Percentage trials by females	—	$r_p(41)$ = .21	1.35	88	—
Mean age	—	$r_p(11)$ = –.20	0.65	75	—
Selected participants	.23 (.41)	$U(22,53)$ = 927.50	1.07	38	—
Unselected participants	.12 (.21)				
Naive participants	.16 (.27)	$t(74)$ = 0.11	0.11	50	—
Experienced participants	.14 (.40)				

APPENDIX, *continued*

		Comparison		Intercoder agreement	
Variable	Mean effect size (*SD*)	Test	*z*	%	ϕ
Receiver sampling characteristics (continued)					
Participants who have studied a mental discipline	.15 (.25)	$t(73) = 0.15$	0.15	100	1.00
Participants who have not studied a mental discipline	.17 (.30)				
All participants students	.06 (.19)	$F(2,75) = 2.87$	1.86	88	—
Some participants students	–.02 (.11)				
No participants students	.21 (.31)				
Sender sampling characteristics [a]					
Percentage trials by females	—	$r_p(19) = .37$	1.79	100	—
Participants are experimenters	.13 (.28)	$t(30) = 0.79$	0.78	88	.77
Participants are not experimenters	.20 (.21)				
All or some senders are friends of receiver	.13 (.29)	$t(20) = 0.72$	0.70	63	—
No senders are friends of receiver	.22 (.24)				

APPENDIX, *continued*

Variable	Mean effect size (*SD*)	Comparison		Inter-coder agreement	
		Test	*z*	%	ϕ
Target selection procedures					
Principal author uses procedure	.14 (.25)	$F(3,34) = 0.14$	0.08	50	—
Other author uses procedure	.22 (.30)				
Other experimenter uses procedure	.17 (.32)				
Person not otherwise involved in study uses procedure	.21 (.24)				
Selection before each trial	.21 (.33)	$t(42) = 1.05$	1.04	75	.49
Selection before whole study	.11 (.26)				
Random number tables	.08 (.20)	$U(27,6) = 501.5$	2.00	75	—
Hardware device	–.03 (.08)				
Repeated measures effects[b]					
No. of trials in study	—	$r_p(30) = -.28$	1.55	63	—
No. of trials per receiver	—	$r_p(27) = -.25$	1.31	88	—
No. of trials per sender	—	$r_p(9) = -.24$	0.71	88	—

APPENDIX, *continued*

Variable	Mean effect size (*SD*)	Comparison		Intercoder agreement	
		Test	*z*	%	ϕ
Receiver procedures					
Experimenter tells receiver success is likely	.16 (.21)	*t*(76) = 0.0	0.00	88	—
Experimenter does not tell receiver success is likely	.16 (.30)				
Receiver undergoes formal relaxation procedure	.09 (.21)	*U*(17,61 = 2469.0	0.72	88	.75
No relaxation procedure	.18 (.31)				
Experimenter in receiver's room during response period	.19 (.33)	*U*(37,15)= 345.0	1.05	100	—
Experimenter not in room	.08 (.14)				
Instructions for mental passivity during response period	.11 (.18)	*U*(13,65)= 466.0	0.64	63	.26
No passivity instructions	.17 (.30)				
Instructions to focus on sender during response period[a]	.30 (.32)	*t*(30) = 2.03	1.95	100	1.00
No orientation toward sender	.10 (.22)				

APPENDIX, *continued*

		Comparison		Intercoder agreement	
Variable	Mean effect size (*SD*)	Test	z	%	φ
Receiver procedures, continued					
Length of response period (for naive receivers only)	—	$r_p(42)$ = .14	0.91	75	—
Response mode includes drawing	.18 (.35)	$t(67)$ = 0.33	0.33	88	—
Response mode verbal only	.16 (.23)				
Mentation report contemporary with response period	.21 (.37)	$t(48)$ = 0.58	0.58	88	—
Mentation report given after response period	.15 (.17)				
Extra mentation details elicited after response period	.41 (.37)	$U(7,60)$ = 1938	2.10	100	1.00
No extra details elicited	.11 (.25)				
Feedback of target identity	.16 (.33)	$t(72)$ = 0.05	0.05	88	.75
No feedback	.17 (.24)				
Receiver knows target location	.08 (.23)	$t(50)$ = 1.95	1.90	50	—
Receiver does not know location	.24 (.38)				

APPENDIX, *continued*

Variable	Mean effect size (*SD*)	Comparison: Test	Comparison: *z*	Intercoder agreement: %	Intercoder agreement: ϕ
Sender procedures[a]					
Instructions to mentally convey target to receiver	.00 (.10)	$U(8,24)$ = 142.0	2.00	75	.47
No such instructions	.20 (.28)				
Instructions to concentrate on target material	.05 (.11)	$U(10,22)$ = 140.5	1.24	88	.77
No such instructions	.20 (.29)				
Sender told receiver's choice of possible target	.08 (.17)	$t(31)$ = 1.85	1.79	38	—
Sender not told receiver's choice	.24 (.32)				
Target type					
Pictures	.05 (.19)	$F(3,66)$ = 6.79	3.29	100	—
People	.26 (.30)				
Geographical sites	.32 (.41)				
Objects	.47 (.34)				

APPENDIX, *continued*

		Comparison		Intercoder agreement	
Variable	Mean effect size (*SD*)	Test	z	%	ϕ
Judging procedures					
Receiver does judging	.07 (.16)	$t(73) = 3.59$	3.43	88	.75
Independent judge does judging	.28 (.34)				
Experienced independent judges	.08 (.24)	$U(12,23) = 170.0$	1.61	88	—
Naive independent judges	.31 (.36)				
No. of items in judging set	—	$r_p(71) = .50$	4.50	88	—

[a] This analysis was conducted only upon telepathy studies in which the author indicated that the person who knew the target identity was a sender, rather than just someone guarding the target or with incidental knowledge of its contents.

[b] Because they were designed to detect practice or boredom effects, these analyses were only applied to studies in which participants did not take part in other trials in the same experiment.