The Journal of Parapsychology, Vol. 58, September 1994

A META-ANALYSIS OF ESP STUDIES CONTRASTING HYPNOSIS AND A COMPARISON CONDITION

BY REX G. STANFORD AND ADAM G. STEIN

ABSTRACT: This meta-analysis examined 25 ESP studies from 12 chief investigators. No potential moderator variable correlated significantly with ESP effect size (π) under either hypnosis or the comparison condition. There was cumulative ESP-test significance for hypnosis, but this significance may be inflated by nonindependence of study outcomes within chief investigator. For both experimental conditions, chief investigator was associated with significant heterogeneity of outcomes. In ESP analyses based on chief investigator, neither the contrast, nor either of the experimental conditions, showed significance. The sum of scored flaws per study did not correlate significantly with outcomes, but the mean number of such flaws was substantial. For the same-subjects studies suitable for analysis, testing order interacted significantly with the experimental manipulation. The hypnosis-comparison contrast was significant only when the comparison condition preceded hypnosis. This significance was due, substantially, to psi-missing in the comparison condition. These and other findings make it difficult to draw substantive conclusions from the current database.

The term *hypnosis*, as used in this paper in reference to experimentation, refers to any experimental condition that includes a hypnotic-induction procedure for the purpose of ascertaining its effect on ESP task performance. That is, *hypnosis* shall be used here, in most cases, as a shorter term for *hypnotic-induction procedure*. This usage is not intended to imply that a particular internal state is necessarily produced, usually or in a specific case, merely by the use of such an induction procedure.

Hypnosis has had a long history of association with putative paranormal events, including ESP (see, e.g., Dingwall, 1967, for an extensive review). Various kinds of folklore, including religious treatises, express the belief that altered states of various kinds favor paranormal phenomena. Not surprisingly, then, the use of hypnotic-induction procedures

The senior author is grateful to St. John's University for a teaching load reduction that aided this work. We gratefully acknowledge helpful suggestions from George Hansen, Victor Labruna, Ephraim I. Schechter, an anonymous referee for the Parapsychological Association 1993 Annual Convention, and an anonymous referee for this journal. We are, grateful to Rhea White, who conducted searches of the PsiLine Database System, and to the Parapsychology Foundation for the use of its facilities.

An earlier version of this paper was presented at the Annual Convention of the Parapsychological Association, August, 1993, Toronto, Ontario, Canada.

Address correspondence to Rex G. Stanford, Psychology Laboratory, SB-15 Marillac Hall, St. John's University, 8000 Utopia Parkway, Jamaica, NY 11439.

has often been seen as a practical, relatively safe, and easy way of trying to enhance extrasensory performance (see individual research reports reviewed herein and reviews of findings, methods, or conceptualization by Honorton, 1977; Honorton & Krippner, 1969; Schechter, 1984; Stanford, 1987, 1992, 1993a).

Schechter (1984) reported a very useful, if preliminary, quantitative analysis of studies involving a contrast of ESP-task performance under hypnosis and comparison conditions. From his statistical examination of findings from studies involving both hypnosis and a comparison (or "control") condition, Schechter concluded that there does appear to be a nonchance difference in hypnosis-comparison ESP-task performance, one favoring hypnosis. He did not, however, provide a definitive statistical analysis to show that performance under hypnosis was significantly different from chance. Stanford (1987) reported such an analysis (which had been done by Schechter at Stanford's request), and it confirmed significant performance in that condition. Stanford (1992) noted, however, that his published estimate of statistical significance for the hypnosis condition might have been inflated because of the probable violation of an assumption underlying the analysis. Specifically, this type of computation assumes independence of outcomes from study to study (Rosenthal, 1991). Two chief investigators contributed four or five studies each, and within-laboratory outcomes might be constrained by situational factors. This would represent a violation of the independence assumption of the statistical analysis, and it might inflate the impression of statistical significance. The present meta-analysis provides investigator-related analyses relevant to this issue.

The goal of the present meta-analytic study was to extend and refine Schechter's work. The Schechter report involved statistical contrasts of the number of studies showing greater ESP-task performance under hypnosis than under a comparison condition and of the number of those studies showing a significant ESP-task difference favoring the hypnosis condition. It did not involve effect-size measures. Its flaws analyses ("Analyses: Stage 2," pp. 7–14) were based on what was effectively a three-level classification of differential success in the hypnosis, as contrasted with the comparison, conditions (i.e., hypnosis significantly higher than comparison; hypnosis nonsignificantly higher than comparison; and comparison nonsignificantly greater than or equal to hypnosis). There were no cases of comparison significantly greater than hypnosis. Those flaws analyses examined just the hypnosis-comparison contrast, not the relationships between flaws and performance under the hypnosis or under the comparison conditions separately. Thus, Schechter's analyses of flaws (1984) had three potential drawbacks: (a) Only a crude measure of differential success (in the hypnosis-

comparison contrast) was used; (b) separate analyses were not undertaken for the hypnosis and comparison conditions; and (c) effect-size measures were not used.

In the present meta-analysis, aggregation of statistical significance will be done separately for the hypnosis and for the comparison conditions. The present analysis will not examine the cumulative significance of the contrast of ESP-task performance for the hypnosis and the comparison conditions. There is a statistical reason for staying away from this contrast in assessing cumulative significance. Of the 25 studies, 21 (84%) used same-subjects designs. Aggregation across studies requires a measure, for each study, of the statistical significance of the contrast. In the present case, we lack in many of the reports (and do not have the data to compute) a measure of the significance of the contrast that takes into consideration that scores under the two conditions may be correlated when derived from the same subjects. The formula (traditionally used in parapsychology) for computing the *z* for the difference of proportions of success under two conditions assumes that the proportions are independent (McNemar, 1969). This assumption is unjustified in the case of same-subjects designs. Given that we are confined here to trial-based analyses because of limitations in the available data, we know of no reasonable alternative statistic in the case at hand to provide the basis of an aggregate measure of the significance of the contrast across conditions. For this reason, we prefer to avoid doing trial-based aggregation for the contrast across a set of studies that includes both within- and between-subjects designs.

However, we will report on how measures of flaws (individually and collectively) and of possible moderator variables correlate with the difference of the ESP-task effect-size measures for the hypnosis and for the comparison conditions. We will also examine, separately for the hypnosis and for the comparison conditions, the correlations of ESP-task effect size with flaws (individually and collectively) and with scores for possible moderator variables.

The present study extended the Schechter analyses by using an effect-size measure (π or pi, the so-called proportion index, Rosenthal & Rubin, 1989) to allow more powerful analyses of the possible role of flaws and of moderator variables, a need suggested in an earlier review (Stanford, 1992). More specifically, the present study, using this more sensitive measure, was intended to investigate the following questions : (a) Under hypnosis, does ESP task performance, cumulated across studies, differ significantly from mean chance expectation? Under comparison conditions? (b) Is ESP-task effect size significantly heterogeneous across studies (for the hypnosis and for the comparison conditions)? (c) Is ESP-task effect size significantly heterogeneous across chief investigators (for the

hypnosis and for the comparison conditions)? (d) Are specific kinds of flaws correlated with ESP-task effect size under hypnosis? Under the comparison condition? With the difference of effect size for the experimental conditions? (e) Is the sum of the flaws, per study, correlated with ESP-task effect size under hypnosis? Under the comparison condition? With the difference of effect size for the experimental conditions? (f) Do potential moderator variables (e.g., the use of specific ESP-enhancement suggestions) correlate with ESP-task effect size for the hypnosis condition? For the comparison condition? With the difference of effect size for the experimental conditions? (g) Were the outcomes of certain studies potentially biased (in favor of the "hypnosis is best" hypothesis) by the number of sessions being allowed to vary across subjects (unbalanced design)? (h) For same-subjects designs do the consequences of testing condition (hypnosis or comparison) depend on the order of testing? That is, does the variable of testing condition interact with order of testing? These were the questions addressed by this meta-analysis.

The present measure of effect size, π, is a nontraditional one developed specifically for situations comparable to those found in ESP research (Rosenthal, 1991; Rosenthal & Rubin, 1989). We regard it as unfortunate that a more traditional measure of effect size, namely, a form of standardized mean difference (see, e.g., Hedges & Olkin, 1985), was unavailable and uncomputable for most of the studies in the present database. Such a measure provides an appealing and easily interpretable way to report a contrast of experimental conditions.

Method

Defining the Database

Kinds of studies to be retrieved. We decided a priori to confine our search to studies that included both hypnosis and comparison conditions, although we knew that there were some studies in the literature that included only the former. This decision was based on three considerations: (a) Our contrast of data for the hypnosis and for the comparison conditions would be most meaningful and less likely to contain confounds if we confined the study to work in which the same laboratory and investigator(s) gathered data for both conditions. For example, within those confines, similar populations of subjects would be maximally likely to have been sampled and to have been studied in comparable environments by the same experimenters. (b) Opening our database to studies including only a hypnosis condition would seem to open our sampling to a greater likelihood of a file-drawer problem based on unpublished,

nonsignificant studies. There is a greater investment of time, effort, and money in a study with both hypnosis and comparison conditions. Therefore, a nonsignificant hypnosis-comparison study might be more likely to be submitted for publication than a nonsignificant hypnosis-only study. Both the author and the editor might see a hypnosis-comparison study as more publishable, regardless of outcome. If they did, such a study would be more likely to be submitted and to be published. A researcher with a nonsignificant hypnosis-only study might consider that its report would be rejected because of lack of a comparison condition. (c) If, as we speculated, significant hypnosis-only studies were considerably more likely to be submitted and published than nonsignificant ones, the inclusion of hypnosis-only work in our database could have favored an inflated estimate of the effect size for hypnosis. These considerations led us to confine our study (as had Schechter, 1984) to studies involving both hypnosis and comparison conditions.

The search. The search for relevant studies was begun by examining the references in the major papers reviewing the hypnosis-ESP literature (Honorton, 1977; Honorton & Krippner, 1969; Schechter, 1984; Stanford, 1987, 1992; Van de Castle, 1969). We continued by making electronic searches of the following databases: (a) American Psychological Association's PsycLIT® on CD-ROM, (b) *Dissertation Abstracts International* on CD-ROM, (c) PsiLine, Templine, Foreign, and NPJP databases in the PsiLine Database System of the Parapsychology Sources of Information Center (PSI, Dix Hills, NY), and (d) a computerized database at the Parapsychology Foundation, Inc., New York (as well as its card catalog). Terms we used in our own searches included *clairvoyance, ESP, extrasensory perception, precognition, telepathy,* and *hypnosis* in various combinations. Rhea White, who searched the PsiLine database, used these terms: *ESP or extrasensory perception or psi, hypnosis,* and *experiment** (which catches *experiments, experimental, experimenter,* etc.), and these combined. Outcomes of these searches will be described in the Results section.

The Effect Size Measure (π) and Its Uses in This Meta-Analysis

The effect size estimator, π, is a proportion index, an index that shows the proportion of correct choices (e.g., for guesses in an ESP test) on a scale in which .50 is mean chance expectation (MCE). This measure of effect size has the specific advantage that it is applicable across studies with ESP tests using varying hit probabilities under the null hypothesis (i.e., differing numbers of target kinds). It converts to a common ground of reference studies involving differing intrinsic (nonpsi) probabilities of a hit. As noted by Rosenthal and Rubin (1989), "When there are more than two equally likely choices, the index π converts the proportion of

hits to the proportion of hits made if there had been only two equally likely choices" (p. 332). Computation of π is described in Rosenthal and Rubin (1989, p. 333). This index, π, was used in almost all of the analyses reported below, including the analyses related to flaws and moderator variables.

Sometimes we wished to compute the z score associated with the value of π observed in a given study. The purpose was to provide a basis for assessing statistical significance in a given study, and, ultimately, such z scores were used in assessing statistical significance across studies (e.g., using the so-called Stouffer method, Rosenthal, 1991). Regrettably, we could compute neither subject-based effect sizes nor subject-based measures of significance for the database as a whole, because we lacked relevant information in a number of studies; subject-based analyses would have been preferable (Stanford & Palmer, 1972). Given these limitations in our database, we used trial-based analyses throughout our meta-analysis, π in the case of effect sizes and z in the case of computing statistical significance. Rosenthal and Rubin (1989, Equation 4, p. 334) provided a formula for computing z based on π and the number of trials in a study. (Their formula assumes large samples of trials.) It is equivalent to (but is computed differently than) the z (normal approximation to the binomial with large samples not requiring continuity correction) that is familiar to parapsychologists. Because we already had π available, we used the Rosenthal-Rubin formula in lieu of the equivalent traditional formula for z that has long been used by parapsychologists.

Classification and Analysis of Flaws

Prior to scoring of flaws we spent numerous hours discussing the nature of flaws in ESP research and how they could, in theory, be classified and scored reliably. Together, we developed an initial set of flaws broader than those used in the analyses reported here. Subsequently, it was discovered that some flaws occurred so infrequently that they would not be useful for this meta-analysis. The flaw categories that were retained were those with utility in the sense just noted and that seemed straightforward in their scoring. We discussed in detail how each flaw category was to be scored. The junior author (A.S.) was responsible for the primary scoring of flaws. Initially, during the process of scoring, he consulted frequently with the senior author (R.G.S.) to ascertain that he understood the agreed-upon classification scheme and how it would apply to the individual case. This showed that, to eliminate ambiguities, some changes were needed in the operational criteria for some of the flaws. A few flaw categories were eliminated because they were either difficult to score or seemed largely redundant with others. Ambiguities

in scoring almost invariably stemmed from a need for refining the criteria used for a given flaw, rather than from lack of clarity in the reports. After the criteria were refined, ambiguities were extremely rare. In those rare cases, both of us discussed the criteria and their applicability to the case at hand. We jointly considered the facts of a report until we reached agreement through looking more closely at the details of the report and assessing their implications. R.G.S. independently made a number of spot checks on the reliability of A.S.'s scoring and found it to concur excellently with his own, once the criteria had been refined. Because of the situational constraints of our project, A.S. was not, strictly speaking, assuredly blind to the outcomes of studies when he scored the flaws for each, but this potential liability was, we feel, well compensated for by the following facts: (a) A.S., while scoring a flaw in a given study, made a conscious effort not to look at (and, in any event, not to scrutinize) the results section of the paper (although the physical paper was handled as an intact entity); and (b) he made efforts to insure that the final criteria for a given flaw were applied uniformly across studies, often going back several times to be certain that this was the case.

The presence of any of the following flaws was scored as "1" and its absence as "0." A flaw was assessed in a study unless something was explicitly said in the report that ruled it out or unless some circumstance was reported that made the presence of the flaw seem unlikely. The scorer made every effort to look for evidence suggesting that a potential flaw was obviated. If no such evidence was found, the flaw was assumed to be present.

Agent and receiver in same room. If, during ESP testing, the receiver was in the same room with the telepathic agent, this flaw was assessed as present. It was assessed as absent when there was no opportunity for sensory influence from an agent (i.e., when the agent was not in the same room with the receiver or when the study did not involve an agent). This variable could thus be scored for all studies, not just those with an agent.

Subject's experimenter may know targets. A flaw was assessed as present if any experimenter in contact with the subject at ESP testing had at any earlier time had sensory contact with information that might have betrayed the identity of one or more of the targets. For example, this flaw was assessed as present if the experimenter shuffled the cards for a given study, even if that was done with the deck underneath a table. This item was scored liberally because (a) the possibilities of subtle sensory communication should not be ignored, and (b) knowledge of even a single target might substantially have affected the results of the study, given the small effect sizes typically found. This flaw was not, however, assessed as present simply because the experimenter who had contact with the sub-

ject(s) had had contact unrelated to specific target sequences with someone who had been involved in target randomization.

Call recorder may know targets. If the experimenter who recorded the subject's calls had potential sensory knowledge of the targets (as indicated by the circumstances discussed under the previous flaw), this flaw was assessed as present. Both this and the previous flaw were assessed as present if an experimenter who had potential sensory knowledge of targets was with the subject during testing and recorded calls. This was because these two flaws seemed to provide two distinct opportunities for compromising the extrasensory nature of the study.

Score accuracy not insured by checking. It seemed extremely unlikely that an investigator who had arranged an independent check on scoring accuracy would have failed to mention it in the report, given that it would have involved an additional, time-consuming, costly, but reassuring, step. It also seemed extremely unlikely that an investigator without an independent check would report the absence of such a check. (We recognize that sometimes an independent check is unnecessary, as when the results are scored by computer. The work under consideration here did not involve computerized testing.) In view of these considerations, this flaw was scored if independent checking was not reported. One referee for this paper suggested that for scoring for this flaw, especially, we should use three categories: present, indeterminate, and absent. The first of these categories would have had no utility because no investigator overtly declared that there had been an absence of independent checking. For reasons already stated, it seemed reasonably clear that a failure to declare this checking meant that it was absent. This seemed the more apparent because an investigator sometimes would make a remark that seemed to indicate sensitivity about this issue, such as saying that the subject was present at the time of checking and observed it. (It seems clear, though, that this is not independent checking.) In our view, two categories are all that are needed or useful for scoring this flaw in the present database. The absence of independent checking in hand-scored parapsychological studies is a potentially serious problem. Both false positives and false negatives can occur. Unintentional false negatives (overlooking hits) might be relatively easy with hand scoring, especially in a condition in which the investigator expects minimal success (e.g., the comparison condition here). In such a case, the record might be checked in a hasty, unmotivated manner that could favor false negatives. There might be particular reason for concern about this in a circumstance in which the comparison condition comes before the hypnosis condition and there is checking of success between these parts of the session. Both investigator and subject might be eager to move on to the exciting hypnosis condition and, so, might proceed with haste. For such

reasons, it seemed important to include this variable and to score it as indicated earlier in this paragraph.

Shuffling instead of random number table. If shuffling, or a functional equivalent, played a fundamental role in target generation, this flaw was assessed as present. A procedure was judged a functional equivalent of shuffling in only one case (Reid, Steggles, & Fehr, 1982). In that study, coin flipping was used in target selection. The shuffling-flaw designation was assessed as absent if a random number table, or a functional equivalent, was used in target generation. Only in one case (Honorton, 1972) was shuffling considered to provide randomization functionally equivalent to what might be provided by a table of random numbers. In that study, for each of the four trials in each session, target selection involved the shuffling of two different decks (25 and 29 cards each) 10 times each and then cutting each deck. The card that appeared on the top of each deck was used, in combination with the equivalent from the other deck, to indicate the target for that trial. The entire procedure was repeated four times for each session to obtain the four targets for the session.

Nonindependent targets across conditions. If the same order or a nonrandom transformation of the target order (e.g., reversed order) was used in the two conditions of a study, this was scored as a flaw.

Design not balanced. If each subject did not have the same number of trials under hypnosis as he or she had under the comparison condition, or if equivalent numbers of runs under the two conditions were different for different subjects, this was scored as a flaw. Either circumstance might inadvertently have biased the data collection so that it favored the hypothesis. Also, an optional-stopping artifact might occur with an unbalanced design of this kind.

Possible Moderator Variables

The presence of any one of the following possible moderator variables in a study was scored as "1." Otherwise, a score of "0" was assigned. (Refer to the Appendix under "Design Variables" for the actual scoring given to each variable in particular studies.

Within-subjects design. This was scored as present whenever the hypnosis-comparison contrast was based on the performance of the same subjects under both conditions.

Induction included test suggestions. This was scored as present if the report indicated that induction involved test suggestions in addition to suggestions for relaxation and sleep. Even a single suggestion of overt behavior (e.g., arm levitation or eye closure) qualified for affirmative scoring on this item. The potential importance of this variable is that overt response to a suggestion serves as a clear cue to the subject (unless

it is perceived as being done by sheer compliance) that he or she is becoming hypnotized and might, therefore, be relevant to expectations that could influence extrasensory outcomes.

Hypnotic suggestions given for success. In 64% of the studies, suggestions were given under hypnosis that were related to having success in, ability for, or confidence concerning the ESP task. This potential moderator variable allows examination of the role of such suggestions.

Subjects selected for hypnotizability. This was scored as present if the report stated that subjects were selected or screened on the basis of any effort at induction, or any testing of overt response to suggestions, or if it was indicated that one or more persons were not included in the study because of failure to pass a test suggestion or "to become hypnotized." No formal standardized test of hypnotic susceptibility was required as the basis of such selection or screening. This item might supply some information relevant to which subjects were genuinely responsive to hypnosis. It might therefore help to indicate whether such responsiveness is a precondition for successful extrasensory response under the hypnosis condition. On the other hand, it is an extremely crude, ill-defined indicator that cannot substitute for testing with standardized instruments for the assessment of hypnotic susceptibility.

ESP task given during hypnosis, not posthypnotically. In the vast majority of studies the ESP test was given during hypnosis. In a small minority (ca. 17%), testing was done with the subject under the influence of a posthypnotic suggestion.

Statistical Analyses

Most of the statistical analyses were performed with SPSS/PC+ (Ver. 4.0). Significant interactions were followed up with a simple-effects program prepared by Dr. Robert Zenhausern, who teaches graduate statistics at St. John's University. Meta-analytic formulae and the formula for the single-mean *t* test were hand-entered, checked for accuracy, and used in a spreadsheet (Lotus).

Analysis of cumulated significance of ESP-task performance by condition (hypnosis or comparison conditions). For a given condition in a given study, the *z* score was computed (across subjects) for ESP-test outcomes. These were then cumulated across studies for the condition in question by the method described by Rosenthal (1991, Equation 4.30, p. 85). The result is itself a *z* score, and the associated probability can be assessed by using a table of the normal distribution. Any conclusion that is to be derived from these analyses must be qualified by the outcomes of the heterogeneity analyses, especially the heterogeneity outcomes related to chief investigator.

Heterogeneity of effect sizes across studies and across chief investigators. A demonstration of heterogeneity of effect sizes across studies (or across

experimenters) can provide justification for examining the roles of flaws and moderator variables. (The effect size for an investigator consists of π based on all the trials for a given condition, across all of that individual's studies.) Significant heterogeneity across studies or investigators suggests that there are limitations on the generalizability of outcomes. Our heterogeneity analyses were based on a formula described in Rosenthal and Rubin (1989, Formula 6a, p. 335). That formula provides a χ^2 statistic with degrees of freedom one less than the number of studies (or investigators). Significant heterogeneity across chief investigators warrants supplementary analyses to learn whether any effect related to hypnosis (or the contrast of comparison-hypnosis) remains significant when investigator π is the basis of the analysis. In the present study, these investigator-based analyses involved *t* tests (see p. 249).

Analysis of flaws. Correlational analyses (r) were used to relate individual flaws and the sum of the flaws to effect-size (π) in the hypnosis and in the comparison conditions and to the difference of π across those conditions. Because these analyses addressed possible threats to the internal validity of these studies, they employed a liberal error (.10, one-tailed) to help insure the detection of such a threat. The one-tailed rejection region varied, for the comparison and the hypnosis conditions, in terms of which end of the distribution allowed potential rejection of the null hypothesis. Specifically, under the hypothesis that one or more flaws are responsible for the "hypnosis is better" finding, a flaw should tend falsely to favor (a) large scores under hypnosis, (b) small scores in the comparison condition, and (c) a difference score favoring hypnosis. Consequently, the rejection region was in the positive tail of the distribution for the hypnosis condition, in the negative tail for the comparison condition, and in the positive tail for the contrast (hypnosis minus comparison).

Although we feel that threats to the preferred interpretation of one's data should be evaluated with a liberal α error such as .10 (lest one misinterpret a finding), some readers might disagree and therefore prefer a different α error for that purpose. Because the value of the inferential statistic will be supplied in every case, each reader is free to set his or her own α error.

Given the importance of flaw analyses, some discussion is warranted about the inferences that are and are not justified from a significant analysis of this kind, regardless of the α error selected. If a flaw analysis is significant, this signals simply that any alleged support for the experimental hypothesis should be viewed cautiously because a flaw conceivably played a role in it. A significant flaw correlation does not, in our view, demonstrate that the flaw explains the primary finding in the database (i.e., that it produced the effect in question). To draw such a conclusion would be to infer causation from correlation. (Sometimes such a conclusion also assumes too much about the percentage of variance accounted for by the correlation.) A significant correlation of a flaw with the dependent variable should, then, be seen as a reason for caution and

as an indicator of a need to further examine the hypothesis in the absence of the flaw.

Some readers might wonder about the appropriateness of reporting Pearson *r*s to assess the relationships between flaws (or moderator variables, as discussed later) and effect sizes, given that the flaw variable (for a single-flaw analysis) takes on only two values. Some might believe we should instead use biserial correlation, or *t* tests. In fact, either approach would lead to precisely the same statistical inference as computing *r*. Several varieties of correlation coefficients discussed in many texts simply represent computationally less complex formulas for computing the Pearson correlation coefficient, *r*, for different types of data (Rosenthal & Rosnow, 1991). Also, *r* can be converted to *t* by means of a simple formula and vice-versa (see Rosenthal, 1991). Computing one instead of the other has no advantage, except that *r* is intuitively easier to interpret for the reader and is itself a popular measure of effect size (unlike *t*). For this reason, we elected to compute and report *r*.

Analysis of potential moderator variables. Correlational analyses (*r*) were used to relate scores of potential moderator variables both to effect size (π) in the hypnosis and in the comparison conditions, and to the difference of π across those conditions.

Assessing potential outcomes bias due to the freedom afforded by allowing an unbalanced design. Correlational analyses (*r*) were used to examine this possibility. Suppose, for example, that an investigator allowed variation in the total number of sessions each subject contributed but administered equal numbers of hypnosis and comparison runs to each subject at each session. If, as seems certain, subjects could discern the experimenter's predictions (i.e., lack of success in the comparison condition, success under hypnosis, and superior performance under hypnosis), those finding themselves not confirming one or more predictions might volunteer for fewer sessions because of feedback from the experimenter about ESP task success. Alternatively, those subjects finding themselves confirming one or more predictions might volunteer for more sessions. Thus, the freedom permitted by this form of unbalanced design might amount to an unintended form of screening for the ability of individual subjects to confirm the experimenter's predictions! This would be a "person confound." It could also introduce an optional-stopping artifact. One way to check on the possibility of this kind of bias in an imbalanced design of the type just discussed would be to correlate, for a given study, the number of sessions completed by each subject with that individual's (a) mean run score in the comparison condition, (b) mean run score in the hypnosis condition, and (c) the difference of the mean run scores under the hypnosis and comparison conditions. If sessions completed showed a substantial positive correlation with (b) or (c) or a negative one with (a), this form of selection bias might have occurred. Because these analyses addressed possible threats to the internal validity of these studies, they employed a liberal α error (.10, one-tailed) in order to help insure the detection of such a threat. Whether the one-tailed rejection

region was in the negative or the positive end of the distribution depended on the considerations just discussed.

Assessing a possible interaction of testing condition with order of testing. The data from three of Casler's studies (1962, main experiment; 1964; 1967) involving within-subjects designs provided an excellent opportunity to learn whether the effects of the experimental manipulation depended on order of testing. These were the only studies in our database that combined (a) information on individual subjects' performance, (b) completely counterbalanced designs, (c) a perfectly balanced design in each study (i.e., equal numbers of trials per condition per subject and equal numbers of trials per subject across conditions), and (d) a fixed number of trials per subject per condition across studies—as well as the same experimenter across the studies. For these reasons, a 2×2 between-within subjects analysis of variance (ANOVA) could be done with conditions (hypnosis-comparison) as the within variable and testing order as the between variable. For this analysis, data could be pooled across the relevant studies, which involved a total of 51 subjects. If the interaction proved to be significant, it was to be followed up with simple-effect analyses and by statistical examination of the mean ESP task performance in each of the cells (combinations of order and testing conditions).

Results and Discussion

Outcomes of Search for Hypnosis-Comparison Studies

We retrieved a total of 29 relevant studies. Of these, 25 (listed in the Appendix) provided information that allowed unequivocal or estimated computation of π. Among the 25, two studies (Honorton, 1972; Reid, Steggles, & Fehr, 1982) provided only approximations of information useful for computing π. Specifically, the Honorton (1972) free-response study did not allow precise computation of π. This was because computation of the raw proportion of hits, which is needed to compute π, might have been compromised by nonindependence of trial outcomes, given that each target also served as a control picture for the other trials. This meant that there well might have been effectively fewer truly independent trials than would have been the case with independence of rating outcomes (Kennedy, 1979). For this reason, π, in the case of Honorton (1972) should be considered an approximation. For the Reid, Steggles, and Fehr (1982) paper, computation of π must also be considered approximate because there seems to be a slight contradiction between the results reported in Tables 1 and 2 of this paper and because the report did not discuss the precise number of hits for the relevant groups. We attempted to resolve these uncertainties but were unable to

contact the senior author. Accordingly, we considered our computation of π for this study to be an approximation. Computations of cumulative significance across studies will be reported that include and that exclude the data of Honorton (1972) and of Reid, Steggles, and Fehr (1982), for the benefit of readers who might object to the inclusion of approximate data. We thought it important to provide computations including these two studies, given that both of them produced outcomes close to MCE in the hypnosis condition. Four additional studies (Casler, 1971; Krippner, 1968a, the latter also published in 1968b; Moss, Paulson, Chang, & Levitt, 1970; Van de Castle & Davis, 1962) could not be included in our meta-analysis database (and, therefore, do not appear in the Appendix) because they did not supply the information that would have allowed computation of π. On the basis of what we know of these four studies, it seems doubtful that their outcomes would substantially modify the conclusions from the present meta-analysis.

All of the 25 studies involved individual testing, and 23 of these involved forced-choice methodology; the other 2 were free-response in character (Braud & Mellen, 1979; Honorton, 1972).

Cumulative Significance of ESP-Task Performance

The cumulative unweighted z score for hypnosis, which is based on the cumulation of performance across all 25 studies (each represented by a z score, Rosenthal, 1991) is significant; $z_H = 8.77$. For the comparison conditions, $z_C = 0.34$, *ns*. If the outcomes of Honorton (1972) and Reid, Steggles, and Fehr (1982) are omitted because π was only approximate for these studies, the result is similar. $z_H = 9.14$, and $z_C = 0.36$. For the entire sample ($N = 25$), the mean of π for the comparison condition is .505, and the standard deviation is .031. For hypnosis, the mean of is .524, and the standard deviation, .035. The significance for hypnosis is probably overestimated, though, because the above computations assume independence across study outcomes, independence that almost certainly does not exist. Also, interpretation of these findings must be qualified in light of the finding that there is striking heterogeneity in this database that is linked to chief investigators.

Heterogeneity of ESP-Task Performance

Across experiments. Heterogeneity for the hypnosis condition across experiments is significant; χ_H^2 (24, $N = 25$) = 121.52, $p < .000001$. Heterogeneity for the comparison condition is also significant across experiments; χ_C^2 (24, $N = 25$) = 45.74, $p < .005$. Excluding the two studies for which π was approximated yields χ_H^2 (22, $N = 23$) = 120.86, $p < .000001$,

and χ_C^2 (22, N = 23) = 45.73, p = .002. There is significant heterogeneity across experiments for both the hypnosis and the comparison conditions, but the degree of such heterogeneity is much greater for hypnosis.

Across chief investigators. Heterogeneity for the hypnosis condition is significant across chief investigators; χ_H^2 (10, N = 11) = 96.98, $p < .000001$. Heterogeneity for the comparison condition is also significant across chief investigators; χ_C^2 (10, N = 11) = 23.02, $p < .011$. Here, too, heterogeneity is much greater for hypnosis. This suggests that some investigators were much better than others at using hypnosis to favor ESP-task success. With regard to the hypnosis condition, comparing the study-based and investigator-based χ^2s (for heterogeneity) in relation to their degrees of freedom suggests that most of the heterogeneity is related to investigator. Figure 1 shows the deviations from MCE for π for each condition for each chief investigator.

A single chief investigator, Reid (Reid, Steggles, & Fehr, 1982), was omitted from Figure 1 and from the heterogeneity computations (above) because he had a single study for which there was no precise way to estimate π. (It seems certain that the inclusion of Reid's work in the heterogeneity analyses would not have changed the conclusion regarding extreme heterogeneity across investigators, especially given that his study produced a result under the hypnosis condition that was apparently close to mean chance expectation—a divergent outcome for this database.)

Investigator-linked heterogeneity made it advisable to examine the effect of the experimental manipulation (comparison-hypnosis) using π for chief investigators as the basis of analysis. A matched-pairs t test based

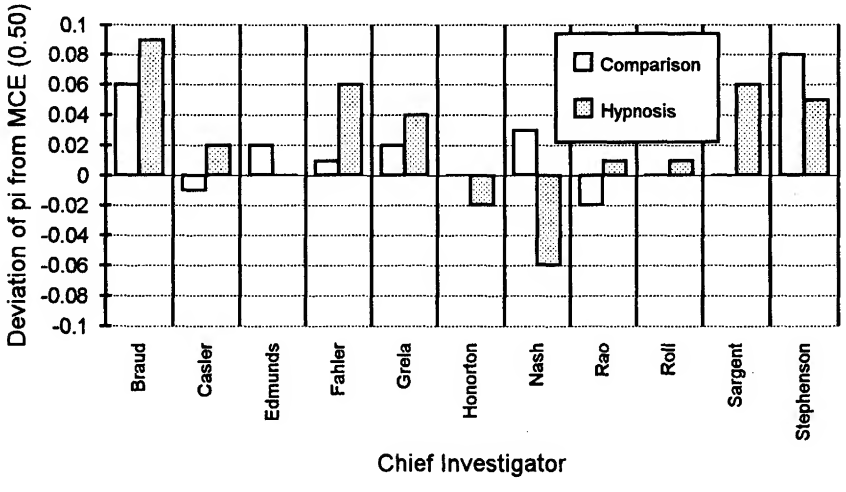

Figure 1. Deviations from MCE (0.50) of π for comparison and hypnosis conditions by chief investigator.

on the 11 chief investigators (see Figure 1) shows that the hypnosis-comparison contrast is not significant; $t(10) = 0.49$, $p = .64$, two-tailed. (Here we can examine the hypnosis-comparison contrast because the contrast is based on π, and our statistic considers the correlation between investigator "performance" under the two conditions. For the present case, $r = .33$.) The mean investigator π for the comparison is .517; $t_c(10) = 1.91$, $p = .085$, two-tailed. For hypnosis, the mean is .524; $t_h = 1.86$, $p = .093$, two-tailed. (Although the mean is smaller in the case of the comparison condition, that condition's outcome has a smaller computed probability under the null hypothesis because of its smaller standard error of the mean.) Thus, analyses that statistically consider the heterogeneity across investigators provide no evidence that hypnosis produces a superior outcome. One caveat is in order. The probabilities associated with these t tests should not be regarded as precise because π will vary somewhat in its reliability across investigators owing to differing numbers of trials from which each is computed. Nonetheless, these analyses do help to clarify the picture of what is—or, more precisely, what is not—happening here. These outcomes do not support the claim that hypnosis per se enhances ESP-task performance. The investigator heterogeneity combined with the analyses reported in this paragraph cast into question the meaning of the overall z scores (Stouffer analyses) presented earlier.

Flaws

Frequency of flaws. The flaws that were examined in relation to π were discussed under Method. Not all possible flaws were examined. Some conceivable flaws did not occur in this database; others occurred so seldom that any analysis of them would have been unreliable. The mean number of flaws per study was 3.40. Table 1 shows the frequency of individual flaws and the proportion of studies in which a given flaw occurred in the total database. Figure 2 shows the frequency of studies with a given total number of flaws.

Flaws and π. Table 2 shows the relevant correlations. There is evidence to suggest that shuffling (or a comparable procedure) in lieu of a random number table (or a comparable procedure) was associated both with inferior performance under the comparison condition and with an enhanced difference favoring hypnosis in the hypnosis-comparison contrast. (Of course, these are not independent findings.)

TABLE 1
HOW FLAWED WERE THE STUDIES? HOW WERE THE STUDIES FLAWED?

Flaw Descriptors	Cases	Proportion
Agent and receiver in same room	3	.12
Subject's experimenter may know target	11	.44
Call recorder may know targets	17	.68
Score accuracy not insured by checking	16	.64
Shuffling instead of random number table	20	.80
Nonindependent targets across conditions	4	.16
Design not balanced	13	.52

These randomization-related correlations remain significant (by our criterion) and do not change substantially if the two studies are excluded ($N = 23$) for which π could only be estimated: r_c (23) = –.32, p = .071, one-tailed; $r_\Delta(23)$ = .31, p = .074, one-tailed.

Some readers might be inclined to dismiss these correlations because they are "significant" only with α set at .10 and are "uncorrected" for the number of flaw analyses done. It should be remembered, however, that these correlations involve a flaw (shuffling) that is relevant to adequacy of randomization of targets. Randomization is one of the most important issues, short of cheating, that can be raised about any ESP database.

TABLE 2
CORRELATIONS OF FLAWS WITH π FOR COMPARISON (πC) AND FOR HYPNOSIS (πH) AND WITH CHANGE IN π ACROSS THESE CONDITIONS ($\Delta\pi$HC) ($N = 25$)

Flaw descriptors	πC	πH	$\Delta\pi$HC
Agent and receiver in same room	.44	.20	–.16
Subject experimenter may know target	.13	–.13	–.21
Call recorder may know targets	–.18	–.21	–.04
Score accuracy not insured by checking	.13	–.02	–.12
Shuffling instead of random number table	–.28[a]	.12	.31[b]
Nonindependent targets across conditions	.12	.09	–.01
Design not balanced	.22	.09	–.09
Total Flaws	.16	.02	–.10

[a] p = .086, one-tailed. [b] p = .064, one-tailed.

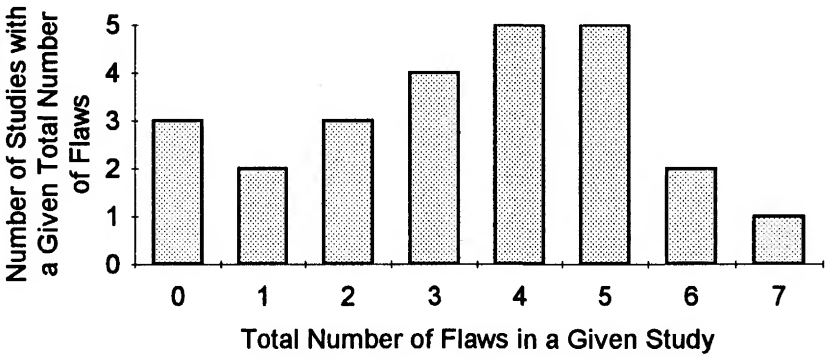

Figure 2. Number of studies with a given total number of flaws.

Proper randomization of targets is essential (a) to obviating nonpsi explanations of various kinds, and (b) to justifying statistical inference (upon which the inference of psi depends). Even so—and even if the reader agrees with us that these correlations should not be dismissed—they should not be over-interpreted. First, neither of them involves the ESP task performance under the hypnosis condition itself. In many respects, the performance there is impressive. (These significant flaw correlations involved only the comparison condition and the contrast of effect sizes.) Second, a correlation of this kind does not prove that the relevant outcomes are caused by randomization problems. This is because causal inferences cannot be sustained from purely correlational data. In our view, these correlations do indicate a need for caution in interpreting ESP performance for the comparison condition and in interpreting the difference in effect sizes for the two conditions (i.e., $\Delta\pi$HC, as in Table 2). (With regard to $\Delta\pi$HC, the reader should note that a given $\Delta\pi$HC [e.g., .054] has an unclear meaning because it is simply a difference of π for the two conditions but does not, per se, reveal which condition provides better evidence of extrachance performance, measured in terms of deviation from MCE. The same can, of course, be said of contrasts of ordinary hitting rates for any two conditions, as Palmer [1975] noted.)

Potential moderator variables and π. Table 3 shows the relevant correlations. Although no potential moderator variable showed statistical significance in these correlations, some are of a magnitude substantial enough to warrant scrutiny in future work. The present database is so small that a sizable correlation (ca. .40, two-tailed) would be required to reach statistical significance. Of special interest for future scrutiny might

be the substantial, potentially counterintuitive trend ($r = -.30$, $p = .143$, two-tailed) suggesting that it may be counterproductive, having induced hypnosis, to give suggestions related to ESP-task success.

Whether ancillary suggestions are counterproductive might depend on whether those suggestions implicitly reinforce or negate the popular belief that the "hypnotic state" per se favors ESP. If a subject believes that hypnosis per se favors ESP, then that belief might be undermined by efforts, during hypnosis, to foster ESP by confidence-building suggestions. Such a strategy might suggest to the subject that the investigator believes hypnosis per se to be insufficient for success and that confidence-building following the induction is necessary. This could frame the whole experience in a different light for the subject. Other kinds of suggestions, given under hypnosis, might actually reinforce subjects' a priori beliefs about hypnosis and ESP. Examples of both types of suggestion are noted below. The implications of this hypothesis cannot adequately be assessed with the current database.

Examples of suggestions that might be counterproductive for such reasons include the potentially ego-involving motivation-related suggestions used by Honorton (1964, 1966, studies that did not show a main effect of the hypnosis manipulation). Those suggestions indicated that subjects had a high degree of ESP ability and were eager to demonstrate this. Subjects who hear such suggestions may think that they have to make ESP happen instead of letting something happen that occurs easily and naturally under hypnosis. This would seem to affront traditional folk beliefs about the special character of hypnosis.

A more productive type of suggestion might be one that seems to flow naturally out of how subjects ordinarily think and feel during hypnosis. That kind of suggestion could reinforce the subject's faith in folk beliefs about hypnosis because it fits naturally into the relatively passive, internal-experience-oriented framework of hypnosis. Examples of this kind include the highly successful hypnotic-dream suggestions used by Braud and Mellen (1979), which were modeled after the very successful hypnotic-dream work of Honorton (e.g., Honorton, 1972; Honorton & Stump, 1969, not reviewed here because it had no comparison condition), and the fruitful suggestions of Sargent (1978) that emphasized relaxation and an internal focus of attention.

The present meta-analytic outcome regarding ancillary suggestions—even if it falls short of statistical significance—underscores the claim of Honorton and Krippner (1969) that suggestions for success are unnecessary for extrasensory success during hypnosis. Experimentation may now be justified on the possibility that some such suggestions can be counterproductive and some productive, depending, perhaps, on whether they imply that some special additional effect or effort beyond

hypnotization is necessary, or whether they fit into the subject's a priori beliefs about what can happen naturally during hypnosis. When hypnosis is administered without explicit suggestions for success and the subject holds traditional folk beliefs about "trance" and ESP, a critical variable may be the subject's belief that he or she is actually hypnotized. The role of belief about having been successfully hypnotized, and its potential interaction with the postinduction suggestions, have been inadequately addressed to date.

TABLE 3

CORRELATIONS OF POTENTIAL MODERATOR VARIABLES WITH π FOR COMPARISON (πC) AND FOR HYPNOSIS (πH) AND WITH CHANGE IN π ACROSS THESE CONDITIONS ($\Delta\pi$HC) ($N = 25$)

Potential moderator variable[a]	πC	πH	$\Delta\pi$HC
Within-subjects design	.09	.09	.09
Induction included test suggestions	.08	−.04	−.05
Hypnotic suggestions given for success	−.05	−.30	−.17
Subjects selected for hypnotizability	.10	−.11	−.13
ESP task given during hypnosis, not post-hypnotically	.33	.09	−.19

[a]No potential moderator variable correlated significantly with any ESP task.

Potential selection bias through unbalanced designs. These analyses were intended to examine the possibility that, given free choice of the number of runs to be done, subjects might have been more likely to continue to participate (or, perhaps, been more socially reinforced for participation by the experimenter) if they were showing trends in line with the transparent hypothesis. Toward this end, we examined studies by Fahler (1957) and by Fahler & Cadoret (1958). These studies provided (a) varied numbers of sessions per subject but (b) equal numbers of runs in comparison and hypnosis conditions at each session. Table 4 reports the correlations between the number of runs done under a given condition by a given subject and the mean hits per run for that condition. Also reported is the correlation between the number of runs done (under either condition) and the hypnosis-comparison difference in the mean hits per run.

TABLE 4
CORRELATIONS OF NUMBER OF RUNS PER SUBJECT WITH MEAN HITS PER RUN FOR COMPARISON AND FOR HYPNOSIS CONDITIONS AND WITH CHANGE IN MEAN HITS PER RUN ACROSS THE HYPNOSIS-COMPARISON CONTRAST (N = NUMBER OF SUBJECTS)

Study	Comparison	Hypnosis	Contrast
Fahler (1957, clairvoyance; $N = 4$)	−.48	.63	.58
Fahler (1957, precognition; $N = 4$)	−.11	.53	.93
Fahler & Cadoret (1958, Series B; $N = 11$)	−.16	−.24	−.08
Fahler & Cadoret (1958, Series C; $N = 12$)	−.26	.72[a]	.73[b]

[a] $p < .0045$, one-tailed. [b] $p < .0035$, one-tailed.

With the exception of Fahler and Cadoret (1958, Series B) there is a trend that reached statistical significance (for the Hypnosis condition and for the contrast) in one series (Fahler & Cadoret, 1958, Series C). The number of runs in which a subject participated related positively to his or her degree of success under hypnosis and to the magnitude of the contrast for the two conditions. All but one series (Fahler & Cadoret, 1958, Series B) showed this trend, albeit not to a significant degree. With sample sizes this small ($N = 4$), the lack of significance of this correlation in certain series is not surprising. The magnitudes of the relevant correlations in the concordant series were strikingly similar, as can be seen from Table 4. Potentially, the freedom to select the number of sessions in which a subject participates can result in a biased selection of data across subjects. Those who are fulfilling the investigator's transparent hypothesis may be more likely to continue. If they do, this would seriously qualify any claimed support for the hypothesis that a hypnotic induction facilitates ESP-task performance across subjects in general. Some subjects may be showing the expected effect, but not others. This kind of freedom might also yield spurious significance as a result of optional stopping. In one of their studies, Fahler and Cadoret (1958, Series B) either did not encounter or managed to obviate this potential problem.

The message here is clear: For testing the hypothesis that hypnosis facilitates ESP task performance, it is important to obviate the potential bias introduced by a failure to set (and equate) in advance the number of sessions and runs done by all subjects.

The interaction of testing conditions and order of testing. Table 5 provides a summary of the ANOVA involving the variables of condition (compari-

son/hypnosis), order, and their interaction. The order effect approached significance. The effect of condition (hypnosis/comparison) was significant, but condition interacted significantly with order. In other words, in the only analyzable block of same-subjects work that we have, the consequences of the hypnosis manipulation, as reflected in the hypnosis-comparison contrast, depended on the order of testing. Figure 3 shows, graphically, the cell means involved in the significant interaction. Table 6 shows descriptive and inferential ESP-test statistics for the comparison and hypnosis conditions, broken down by order. The most striking ESP performance in any of the cells of this ANOVA design was the psi-missing when the comparison condition preceded hypnosis.

TABLE 5

ANOVA FOR DATA FROM CASLER, 1962 (MAIN EXPERIMENT), 1964, 1967: CONDITIONS (WITHIN), ORDER (BETWEEN), AND THEIR INTERACTION

Source of Variation	*SS*	*df*	*MS*	*F*	*p*
Between subjects					
Order	44.55	1	44.55	2.88	.096
Subjects within order	759.04	49	15.49		
Within subjects					
Condition (comparison vs. hypnosis)	183.86	1	183.86	19.15	.000
Order × condition	41.19	1	41.49	4.29	.044
(Condition × subjects.) within order	470.40	49	9.60		

To further examine the condition × order interaction, we conducted simple-effects analyses. The order effect for the comparison condition approached significance; $F(1, 49) = 3.42$, $p = .07$. There was no hint of an order effect for the hypnosis condition; $F(1,49) = 0.00$. The effect of the hypnosis-comparison manipulation was not significant when hypnosis was administered first; $F(1, 49) = 2.66$, $p = .11$. The hypnosis-comparison manipulation was significant only when hypnosis was administered second; $F(1, 49) = 20.79$, $p = .0003$. Figure 4 illustrates, for the three studies separately, this hypnosis-comparison contrast when hypnosis came second. It indicates the consistency of the effect.

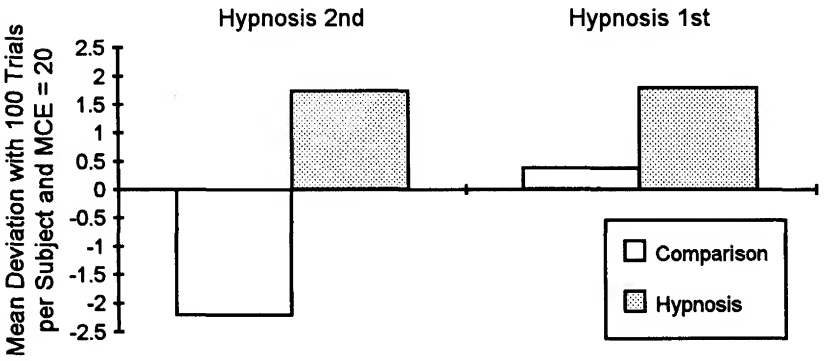

Figure 3. How testing order affects results of the hypnosis manipulation. Results are based on Casler, 1962 (Main Experiment), 1964, 1967.

TABLE 6
DESCRIPTIVE AND INFERENTIAL STATISTICS FOR CONDITIONS EXAMINED BY ORDER (FOR DATA FROM CASLER, 1962, MAIN EXPERIMENT; 1964; 1967; 100 TRIALS PER SUBJECT AND MCE = 20)

Condition	Hypnosis 2nd	Hypnosis 1st
Comparison	$M = 17.78$	$M = 20.38$
	$n = 27$	$n = 24$
	$SD = 3.53$	$SD = 3.96$
	$t = -3.27$	$t = 0.46$
	$p = .003$	$p = .65$
Hypnosis	$M = 21.74$	$M = 21.79$
	$n = 27$	$n = 24$
	$SD = 3.45$	$SD = 2.86$
	$t = 2.62$	$t = 3.07$
	$p = .014$	$p = .005$

These findings suggest that there may be asymmetry of transfer in moving from comparison to hypnosis (and vice versa) in same-subjects designs in this domain. There was a striking tendency to psi-miss in the comparison condition when it preceded hypnosis, but not when it followed hypnosis. One possible interpretation of this finding is that

subjects are annoyed at having to sit through a comparison condition when what they really want is the excitement of hypnosis. Alternatively, the need for suppressing performance under the comparison condition may be particularly salient when subjects are eagerly awaiting the hypnosis condition. Other interpretations are possible. In any event, the psychological meaning of the comparison condition may be very different for subjects who experience that condition before hypnosis and for those who experience it afterward, as judged from the work of Casler. Realistic interpretation of the results of same-subjects designs in this domain requires examination of a possible interaction between the hypnosis manipulation and the order of testing. Alternatively, one could use independent-groups designs to obviate these and other problems that can inhere in same-subjects designs (see Stanford, 1987). Interactions of hypnosis with testing order should, though, be regarded as more than just nuisance effects. They may have psychological meaning worthy of investigation.

The major raw data of this meta-analysis are provided in the Appendix.

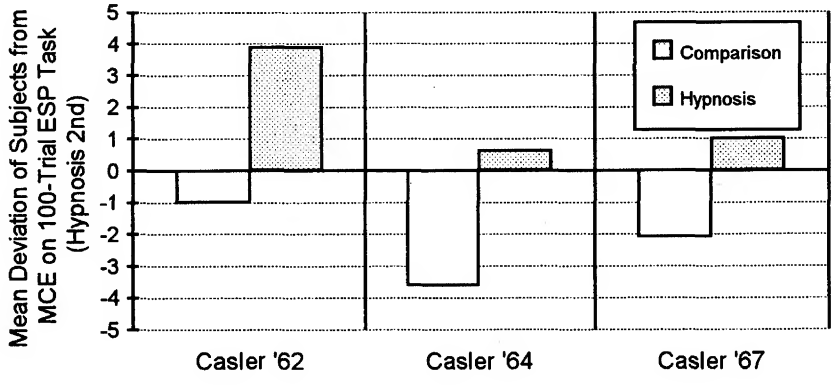

Figure 4. Breakdown by studies of consequences of the hypnosis manipulation when hypnosis comes second. Results are based on Casler, 1962 (Main Experiment), 1964, 1967.

Concerning Matters Not Considered

In our analyses there was no consideration of the problem of nonrandom assignment to experimental conditions in between-subjects

designs. Statistical analyses of the possible consequences of this problem would depend on a far larger sample of between-subjects designs than were at hand. Also, this problem occurred in one degree or another (or had potential for having occurred) in all four of the studies in the meta-analysis that involved between-subjects designs. (It also occurred in the not-included Moss et al., 1970, study). For these reasons, there was no possibility for analyses that might have illuminated any role played by this flaw.

Nor was it possible to do an analysis of the role of experimenter effects on the basis of studies that did and did not provide protection against this potential problem. This was because in every study we retrieved it would appear that the experimenter who worked with the subject during ESP testing knew whether the subject was hypnotized (or had been, in the case of posthypnotic suggestion).

We did not do an analysis of a potential problem of nonindependent targets across subjects—sometimes called a "stacking effect"—because no studies involved in this meta-analysis had this problem. Similarly, we did no analysis of the possible problem of subjects having direct sensory contact with the targets, because this problem did not appear to exist in any of the studies.

No moderator-variable analysis was reported with the design variable of "Number of alternatives in forced-choice study ('K' in Rosenthal's notation)" as listed in the Appendix. Preliminary moderator-variable analysis showed that although this variable was producing significant outcomes, this was due to (a) the small number of studies with $K \neq 5$ combined with (b) the effect of extreme values (i.e., $K = 20$ and $K = 10$) in two studies, little variability outside the two extreme cases, and the small total sample of studies. Under these circumstances, reporting the outcome would have been misleading.

Some readers may be puzzled that no analysis has been reported here to try to rule out a file-drawer problem in this database. None has been reported because it would probably be misleading in the degree to which it would suggest that the data are not threatened by such a problem. Stanford (1992) has discussed two reasons for skepticism about the typical approach to the file-drawer problem as it has been applied and interpreted in many cases. The traditional approach involves the computation of the number of unreported or unretrieved studies averaging null results that would be needed to wipe out the statistical significance of the retrieved database. One of our reasons for skepticism is that the relevant computations assume independence of study outcomes (Rosenthal, 1991, p. 105), but this is an unreasonable assumption for the present database. Computation of a file-drawer analysis might be especially misleading in this case because specific investigators have a history

of outstanding success in this area, have contributed several studies, and have contributed results that appear to weigh disproportionately in the cumulative outcome. The results of our analyses of heterogeneity of investigator outcomes, especially in the hypnosis condition, are compatible with this impression. Also compatible is the fact that the heterogeneity in our database appears to depend more on investigators than on studies, as suggested by the outcomes of the heterogeneity analyses. In the hypnosis condition, the 11 investigators produced statistical evidence of heterogeneity that was nearly as significant as that produced by the 25 studies. (Insufficient studies by several investigators made it impossible, though, to do the kind of analysis of the nonindependent-outcomes hypothesis that would optimally address this problem.) It seemed prudent, then, not to do a "fail-safe" analysis on the present database. Such an analysis also makes little sense in light of the finding that the data failed to indicate a significant effect of hypnosis when effect size of investigator was the basis of the analysis. As was noted in the Results section, this likely nonindependence of outcomes means that the Stouffer analyses might have overestimated the statistical significance of the results in the hypnosis condition.

Conclusions

Although the results with the Stouffer (cumulative significance) analyses look very promising for the hypothesis that hypnosis favors ESP task performance, the other analyses reported here raise questions about how this positive finding should be interpreted. Results are strikingly heterogeneous across chief investigators, especially in the hypnosis condition. Accordingly, when the effect size associated with a given chief investigator is considered as the basis for statistical analysis, the hypnosis-comparison contrast does not approach statistical significance. Instead, performance levels in the comparison and the hypnosis conditions are similar, with a trend toward psi-hitting. Whatever the role of hypnosis might be, the success of using that approach is very uneven across investigators.

In some ways this is to be expected. The investigator who is to use hypnosis effectively must have strong interpersonal skills and a good knowledge of hypnosis. It is unlikely that all investigators are equal in these regards. Presumably, the success of hypnosis also depends somehow on the characteristics of those used as hypnotic subjects. Different investigators might have had access to populations that differed in the skills and personal attributes necessary for successful hypnosis. Also, the success of the technique might depend in some degree on the

expectations of the investigator, and these studies lacked controls for such effects (as noted, for example, by Schechter, 1984). For these reasons, the heterogeneity of effects across investigators should not be taken to imply that hypnosis has no utility for enhancing extrasensory performance. It is simply that the extant studies provide more puzzles than useful clues about the boundary conditions for success in this domain.

George Hansen (personal communication) has remarked about the strikingly high percentage of studies in this meta-analysis that had 4 or fewer subjects (8 of 25 studies, 32%). In our view, this is another reason for caution about concluding that the hypnosis findings generalize across subjects.

Hypnosis may very well be useful, when combined with currently unspecifiable circumstances, in fostering ESP-task success. Enough investigators had success (and enough did not) to suggest this possibility. The reasons for the differential success are unclear. A major problem is the lack of systematic study of potentially relevant variables such as those discussed in the next-to-last paragraph. We, for example, had insufficient data to assess the role of hypnotic susceptibility, as measured by standardized tests, in ESP-task performance under hypnosis (and under the comparison condition).

Expectations, including those transmitted from the experimenter to the subject, may be important to parapsychological outcomes, as they are in many other aspects of work with hypnosis. Kirsch (1991), in the description of his social learning theory of hypnosis, has discussed the role of expectancies in nonparapsychological studies of hypnosis. We lacked sufficient data in hypnosis-ESP studies to assess the role of expectations.

Several investigators who did more than one hypnosis-ESP study consistently found that hypnosis seemed useful in eliciting ESP-task success. Honorton, another such investigator, very consistently found that hypnosis was not successful in eliciting ESP task success across subjects in general. Honorton, from the earliest stages of his career (1964, and see Stanford, 1993b, for a review) studied the very reasonable hypothesis that whether hypnosis is successful in enhancing ESP performance depends on the characteristics of the individual tested. His three studies provided no evidence that the use of hypnotic induction per se can enhance ESP performance. (He did find evidence, though, that it might do so with suitable subjects, as determined by an external criterion.) His work always seemed to point toward a person × situation interaction, rather than a main effect of the technique itself. Present knowledge does not provide an understanding of these paradoxical outcomes. The truth may be hidden somewhere in a mix of investigator expectancies,

characteristics of subjects tested, skills of investigators as hypnotists, and the content of postinduction suggestions.

Presently, we do not even know what subject characteristics, if any, might mediate (or, perhaps, moderate) any effect of the hypnosis manipulation. Methodologically and conceptually improved work is needed that will systematically explore the bases of success and failure using this method.

Thoughts on Assessing the Role of Flaws in Meta-Analyses

Although the mean number of scored flaws per study was 3.4, the total-flaws measure did not correlate significantly with outcomes for the comparison, for hypnosis, or for the hypnosis-comparison contrast. These facts provide some reassurance, despite the substantial mean number of scored flaws.

On the other hand, a serious effort at interpreting this database should not ignore the high rate of flaws, despite the null results for the correlation with total flaws. It is sometimes suggested that null correlations such as this show that there is no reason to worry about the internal validity (in this case, the extrasensory character) of the studies having been compromised. Reality may not be this simple, especially with flaws as prevalent as they were in this database. There are at least four reasons why the present null, total-flaws analysis should not foster complacency about possible artifact(s) in this database. *These same considerations presumably apply to all meta-analyses, whether inside or outside of parapsychology, in which the investigators have discounted the effects of flaws because of a null, unweighted total-flaws correlation.*

1. The flaws were scored simply as present or absent in this analysis even though many of the flaws, if present, might vary in magnitude. For example, it is one thing to say that the experimenter had potential knowledge of one or more of the targets, but it is another to estimate how many he or she might have known. Therefore, analyses such as ours (including those in numerous other meta-analyses) are relatively crude and may lack the sensitivity needed fully to assess the role of potential artifacts. Doing flaws ratings of a more sensitive type would have involved greater subjectivity of judgment, extraordinarily elaborate sets of rules to try to obviate that subjectivity, and far more work and time than would have been justified on a database this small. The best we can do here is to forewarn the reader that total-flaws analyses such as this are relatively crude and may lack the sensitivity needed to disclose any role for artifacts.

2. The flaws were weighted equally in this analysis, but they might not be equally important in compromising the extrasensory interpretation of the data. (This is one reason that correlations involving individual flaws were reported. The effect of an important flaw might have been swamped in the total-flaws analysis.) An unweighted total-flaws analysis might obscure more than it reveals. In a weighted-flaws analysis, the weights assigned to individual flaws would vary depending on the a priori assumptions of the individual analyst about the importance of a given flaw. In the present meta-analysis, instead of doing a weighted-flaws analysis reflecting our own, perhaps idiosyncratic, presuppositions, we opted for an unweighted analysis and provided the raw data. Others can weight the flaws as they wish and draw their own conclusions. (It is particularly important that if weighting be done, it be done blind to the outcomes of the studies. Because we did not plan a weighted-flaws analysis in advance of having examined the relevant data, we could not do a weighted-flaws analysis free of potential bias because of our knowledge of outcomes.)

3. If a meta-analyst is interested in assessing the possible role of a particular kind of artifact in a database, not all kinds of flaws are relevant to the operation of that artifact. Some are more relevant than others, and sometimes the effects of a particular flaw might logically depend on the magnitude (or simply the presence) of another flaw. For example, if the artifact of concern is "sensory communication," the consequences of having had the agent and receiver in the same room might interact with a flaw such as nonindependent targets across conditions. Nonindependence of targets across conditions might exacerbate the untoward consequences of any sensory leakage between agent and receiver. Some of the consequences of this randomization flaw, then, would be nonadditive because they would depend on sensory leakage. Simply weighting individual flaws according to importance (or even just counting them) and then summing them suggests that each has only an additive (independent) effect in favoring an artifact, but that might be an unreasonable assumption. If a meta-analyst wishes to examine a hypothesis about a particular kind of artifact, he or she perhaps should decide what potential flaws favor that artifact and whether their effects should logically be additive, interactive, or both. Only after this step has been taken can a suitable index (call it "artifact vector") for that particular artifact be computed. This approach involves creating a model for how a particular flaw functions in relation to a specific kind of artifact and then using that model to compute the vector for that artifact. We suggest that model building is an important direction for future flaws analyses. In the case of the present database, though, this approach would seem to be overkill. The present database was simply too small, in our view, to sustain this type of detailed analysis. (For example, only a single study combined the flaws of "agent and receiver in same room" and "nonindependent targets across conditions," although three studies had the former flaw, and four had the latter.) In a small database, combining flaws can result in a

sample of concordant cases that is so small that the results are unreliable and, hence, potentially misleading. Having thus exculpated ourselves for not having done these more sophisticated kinds of model-based analyses, we must remind the reader that our total-flaws analysis examined only an additive model for unweighted flaws.

4. The reported total-flaws analysis, by its very nature, looked only at the possibility of a linear relationship between the number of individual flaws and ESP-task success. Among the several possibilities for nonlinearity would be no effect of sum of flaws (or of magnitude of a given flaw in the case of single-flaw analyses using scaled flaw measures) until that variable reaches a particular magnitude, after which it has a maximal effect. Traditional correlation would be relatively insensitive to a threshold effect such as this.

The following conclusions, then, apply to the topic of flaws: The fact that the null hypothesis was not rejected by the total-flaws analysis does not rule out the possibility of flaws as sources of artifact in this database. The present considerably flawed sample of studies is too small to support more sophisticated flaws analyses. When and if this database is substantially expanded by additional, improved work, more extensive and suitable analyses would be possible. On the other hand, the present inability to rule out all feasible flaw models cannot reasonably be seen as a basis for dismissing as artifactual the ESP-task results of the present database. Positive evidence would be needed to support such a conclusion.

In the present database, potential inadequacy of randomization, due to shuffling, was, by our criterion, inversely related to success in the comparison condition, but shuffling was not reliably correlated with success under hypnosis. Consequently, any problems related to shuffling have not been shown to threaten the parapsychological interpretation of the results under hypnosis. The same cannot be said, though, of the hypnosis-comparison contrast because it is affected by the shuffling-success relationship observed for the comparison condition. Although caution is warranted in interpreting outcomes for the comparison condition and for the contrast, it should not be inferred that a shuffling flaw definitely affected the outcomes for those two measures. Correlation cannot be the basis of causal inference. In the present case, the correlation signals a reason for caution and a need for improved methodology.

In our opinion, what is known of the present database provides substantial justification for future work with improved methodology. Only such work can provide confidence about what happens when investigators use hypnosis to try to facilitate ESP-task performance.

Design-Related Caveats

Most of the studies used same-subjects designs, and we have shown that order of testing can interact with the manipulation (hypnosis-comparison), thereby qualifying the interpretation of any main effect of the manipulation. Future work can profit either by avoiding same-subjects designs or by checking on and reporting information about interactions of conditions and order.

Some of the studies that used same-subjects designs might inadvertently have provided an opportunity for those who were doing better under hypnosis to be tested more. This is because certain of the same-subjects studies allowed variations in the amount of testing across subjects. There was evidence suggesting that, in some such work, subjects were tested for more sessions whose outcomes tended more strongly to support the hypothesis of hypnotic facilitation of ESP-task performance. The circumstances that allow this possibility should be avoided if studies are to support inferences about the effects of hypnotic induction.

Prospect

Although the findings of this meta-analysis seriously complicate efforts to interpret the present database, they point toward the importance of new, methodologically improved, work in this intriguing and promising area. Future work in this domain might profit by attentiveness to the concerns raised in this meta-analysis and by the suggestions for hypnosis-ESP research provided in earlier reviews of this literature (Honorton & Krippner, 1969; Schechter, 1984; Stanford, 1987, 1992, 1993a).

APPENDIX:
HYPNOSIS-ESP META-ANALYSIS DATA

Studies (N = 25)	Flaws								Design Variables					
Author(s) and Date	I	II	III	IV	V	VI	VII	Sum	A	B	C	D	E	F
Braud & Mellen '79	0	0	0	1	0	0	0	1	1	1	1	0	1	2
Casler '62 Preliminary Experiment	1	1	1	1	1	0	1	6	0	1	1	0	1	5
Casler '62 Main Experiment	0	0	1	1	1	0	0	3	1	1	1	0	1	5
Casler '64	0	0	1	1	1	1	0	4	1	1	1	0	1	5
Casler '67	0	0	1	1	1	1	0	4	1	1	0	1	1	5
Casler '76 Group 1	0	0	1	1	1	0	0	3	1	1	1	1	1	5
Edmunds & Jolliffe '65	0	0	1	1	1	0	1	4	1	1	1	1	1	5
Fahler '57	0	1	1	0	1	1	1	5	1	0	0	1	1	5
Fahler & Cadoret '58 Section A	1	1	1	1	1	0	1	6	1	0	0	0	1	5
Fahler & Cadoret '58 Section B	0	1	1	0	1	0	1	4	1	0	0	0	1	5
Fahler & Cadoret 58 Section C	0	0	0	0	1	0	1	2	1	0	0	0	1	5
Grela '45	0	0	0	1	1	0	1	3	1	1	1	1	0	5
Honorton '64	0	1	1	1	1	0	1	5	1	1	1	1	1	5
Honorton '66	0	0	0	0	0	0	0	0	1	1	1	1	1	5
Honorton '72	0	0	0	0	0	0	0	0	0	1	1	0	1	4
Nash & Durkin '59	0	1	1	0	0	0	0	2	1	0	1	0	1	10
Rao '64	0	1	1	0	1	0	0	3	1	1	1	1	0	5
Rao '79 Series 1	0	0	1	1	1	0	1	4	1	0	1	0	0	5
Rao '79 Series 2	0	0	0	0	1	0	0	1	1	0	1	0	0	5
Reid, Steggles, & Fehr '82	0	0	0	1	1	0	0	2	0	0	0	0	1	2
Roll '75 Series I	0	1	1	1	1	0	1	5	1	0	0	0	1	5
Roll '75 Series II	0	1	1	1	1	0	1	5	1	0	0	0	1	5
Roll '75 Series III	0	1	1	1	1	0	1	5	1	0	0	0	1	5
Sargent '78	0	0	0	0	0	0	0	0	0	0	1	0	1	5
Stephenson '65	1	1	1	1	1	1	1	7	1	1	1	1	1	20

Flaw Codes (see text)

I. Agent and receiver in same room
II. Subject's experimenter may know targets
III. Call recorder may know targets
IV. Score accuracy not insured by checking
V. Shuffling instead of random number table
VI. Nonindependent targets across conditions
VII. Design not balanced

Design Variables (see text)

A. Within-subjects design
B. Induction included test suggestions
C. Hypnotic suggestions given for success
D. Subjects selected for hypnotizability
E. ESP task given during hypnosis, not post-hypnotically
F. Number of alternatives in forced-choice study ("K" in Rosenthal's notation)

	Comparison				Hypnosis				Contrast
Authors	Ss	Trials	Hits	π	Ss	Trials	Hits	π	Δπ
B & M '79	10	100	56	0.560	10	100	59	0.590	0.030
C '62 p.e.	26	5200	1023[a]	0.495	22	4400	916	0.513	0.018
C '62 m. e.	10	1000	175	0.459	10	1000	222	0.533	0.074
C '64	7	700	140	0.500	7	700	150	0.522	0.022
C '67	7	700	152	0.526	7	700	161	0.544	0.018
C '76 Grp 1	10	1000	197	0.495	10	1000	212	0.518	0.023
E & J '65	4	1600	340	0.519	4	8000	1594	0.499	–0.020
F '57	4	4500	894	0.498	4	4500	1003	0.534	0.036
F & C '58 A	*	2825	644	0.542	*	3700	916	0.568	0.027
F & C '58 B	11	2625	531	0.504	11	2625	643	0.565	0.061
F & C '58 C	12	3000	599	0.499	12	3000	783	0.586	0.086
G '45	11	2375	513	0.524	11	1975	444	0.537	0.013
H '64	6	1100	214	0.491	6	1100	193	0.460	–0.032
H '66	20	2500	510	0.506	20	2500	487	0.492	–0.014
H '72	30	120	30	0.500	30	120	30	0.500	0.000
N & D '59	2	600	67	0.531	2	600	48	0.439	–0.092
R '64	1	500	82	0.440	1	500	113	0.539	0.099
R '79 1	1	800	144	0.468	1	600	120	0.500	0.032
R '79 2	20	4000	784	0.494	20	4000	833	0.513	0.019
R, S, & F '82	20	120	60	0.500	20	120	60	0.500	0.000
R '75 I	1	750	169	0.538	1	1125	241	0.522	–0.016
R '75 II	1	750	142	0.483	1	1400	294	0.515	0.032
R '75 III	1	750	138	0.474	1	1125	228	0.504	0.030
S '78	20	1000	201	0.502	20	1000	238	0.555	0.054
S '65	25	675	46	0.582	25	1160	69	0.546	–0.036

[a]The number of hits here has been adjusted on the basis of a correction reported later (Casler, 1982)

*Relevant data were not reported.

References[1]

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

*Casler, L. (1962). The improvement of clairvoyance scores by means of hypnotic suggestion. *Journal of Parapsychology*, **26**, 77–87.

*Casler, L. (1964). The effects of hypnosis on ESP. *Journal of Parapsychology*, **28**, 126–134.

*Casler, L. (1967). Self-generated hypnotic suggestions and clairvoyance. *International Journal of Parapsychology*, **9**, 125–128.Casler, L. (1971). Hypnotically induced interpersonal relationships and their influence on GESP. *Proceedings of the Parapsychological Association 1969*, **6**, 14–15. (Abstract)

[1]Note: * denotes reports included in the database for the meta-analysis.

*Casler, L. (1976). Hypnotic maximization of ESP motivation. *Journal of Parapsychology*, **40**, 187–193.

*Casler, L. (1982). Correspondence. *Journal of Parapsychology*, **46**, 289–290.

Dingwall, E. J. (Ed.). (1967). *Abnormal hypnotic phenomena* (Vols. 1–4). London: Churchill.

*Edmunds, S., & Joliffe, D. (1965). A GESP experiment with four hypnotized subjects. *Journal of the Society for Psychical Research*, **43**, 192–194.

*Fahler, J. (1957). ESP card tests with and without hypnosis. *Journal of Parapsychology*, **21**, 179–185.

*Fahler, J., & Cadoret, R. J. (1958). ESP card tests with and without hypnosis. *Journal of Parapsychology*, **22**, 125–136.

*Grela, J. J. (1945). Effect on ESP scoring of hypnotically induced attitudes. *Journal of Parapsychology*, **9**, 194–202.

*Honorton, C. (1964). Separation of high- and low-scoring ESP subjects through hypnotic preparation. *Journal of Parapsychology*, **28**, 250–257.

*Honorton, C. (1966). A further separation of high- and low-scoring subjects through hypnotic preparation. *Journal of Parapsychology*, **30**, 172–183.

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

Honorton, C. (1977). Psi and internal attention states. In B. B. Wolman (Ed.), *Handbook of parapsychology* (pp. 435–472). New York: Van Nostrand Reinhold.

Honorton, C., & Krippner, S. (1969). Hypnosis and ESP performance: A review of the experimental literature. *Journal of the American Society for Psychical Research*, **63**, 214–252.

Honorton, C., & Stump, J. (1969). A preliminary study of hypnotically-induced clairvoyant dreams. *Journal of the American Society for Psychical Research*, **63**, 175–184.

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

Kirsch, I. (1991). The social learning theory of hypnosis. In S. J. Lynn & J. W. Rhue (Eds.), *Theories of hypnosis: Currents models and perspectives* (pp. 439–465). New York: The Guilford Press.

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

Krippner, S. (1968b). An experimental study in hypnosis and telepathy. *The American Journal of Clinical Hypnosis*, **11**, 45–54.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Moss, T., Paulson, M. J., Chang, A. F., & Levitt, M. (1970). Hypnosis and ESP: A controlled experiment. *The American Journal of Clinical Hypnosis*, **13**, 46–56.

*Nash, C. B., & Durkin, M. G. (1959). Terminal salience with multiple digit targets. *Journal of Parapsychology*, **23**, 49–53.

Palmer, J. (1975). Three models of psi test performance. *Journal of the American Society for Psychical Research*, **69**, 333–339.

*Rao, K. R. (1964). The differential response in three new situations. *Journal of Parapsychology*, **28**, 81–92.

*Rao, K. R. (1979). Language ESP tests under normal and relaxed conditions. *Journal of Parapsychology*, **43**, 1–16.

*Reid, G., Steggles, S., & Fehr, R. C. (1982). State, emotionality, belief, and absorption in ESP scoring. *Journal of The Association for the Study of Perception*, **17**, 28–39.

*Roll, W. G. (1975). *Theory and experiment in psychical research.* New York: Arno. (See, in particular, pp. 248–271.)

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, **106**, 332–337.

*Sargent, C. L. (1978). Hypnosis as a psi-conducive state: A controlled replication study. *Journal of Parapsychology*, **42**, 257–275.

Schechter, E. I. (1984). Hypnotic induction vs. control conditions: Illustrating an approach to the evaluation of replicability in parapsychological data. *Journal of the American Society for Psychical Research*, **78**, 1–27.

Stanford, R. G. (1981). Are we shamans or scientists? *Journal of The American Society for Psychical Research*, **75**, 61–70.

Stanford, R. G. (1987). Ganzfeld and hypnotic-induction procedures in ESP research: Toward understanding their success. In S. Krippner (Ed.), *Advances in parapsychological research 5* (pp. 39–76). Jefferson, NC: McFarland.

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

Stanford, R. G. (1993a). ESP research and internal attention states: Sharpening the tools of the trade. In L. Coly & J. D. S. McMahon (Eds.), *Psi research methodology: A re-examination,* (pp. 189–242). New York: Parapsychology Foundation.

Stanford, R. G. (1993b). Learning to lure the rabbit: Charles Honorton's process-relevant ESP research. *Journal of Parapsychology*, **57**, 129–175.

Stanford, R. G., & Palmer, J. P. (1972). Some statistical considerations concerning process-oriented research in parapsychology. *Journal of The American Society for Psychical Research*, **66**, 166–179.

*Stephenson, C. J. (1965). Cambridge ESP-hypnosis experiments, 1958–64. *Journal of the Society for Psychical Research*, **43**, 77–91.

Van de Castle, R. L. (1969). The facilitation of ESP scores through hypnosis. *The American Journal of Clinical Hypnosis*, **12**, 37–56.

Van de Castle, R. L., & Davis, K. R. (1962). The relationship of suggestibility to ESP scoring level (abstract). *Journal of Parapsychology*, **26**, 270–271.

Psychology Laboratory
SB-15 Marillac Hall
St. John's University
8000 Utopia Parkway
Jamaica, NY, 11439